Patterns, Functions and Structures on a protein topology database

www.tops.leeds.ac.uk

D. Gilbert, A.C. Tan, G. Torrence, M. Veeramalai, J. Viksna

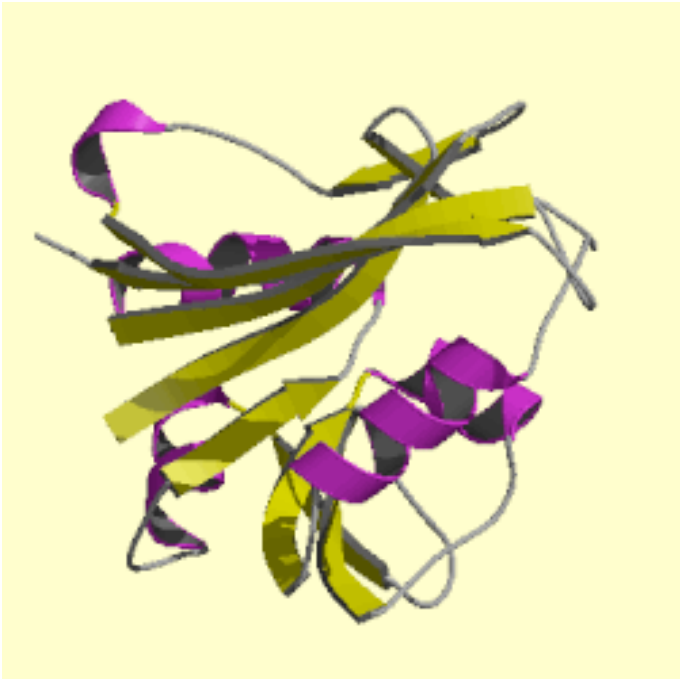
Bioinformatics Research Centre
Department of Computing Science, University of Glasgow

D. Westhead, I. Michalopoulos

Department of Biochemistry & Molecular Biology
University of Leeds

Structure and function?

Dihydrofoliate reductase



Human (1drf)

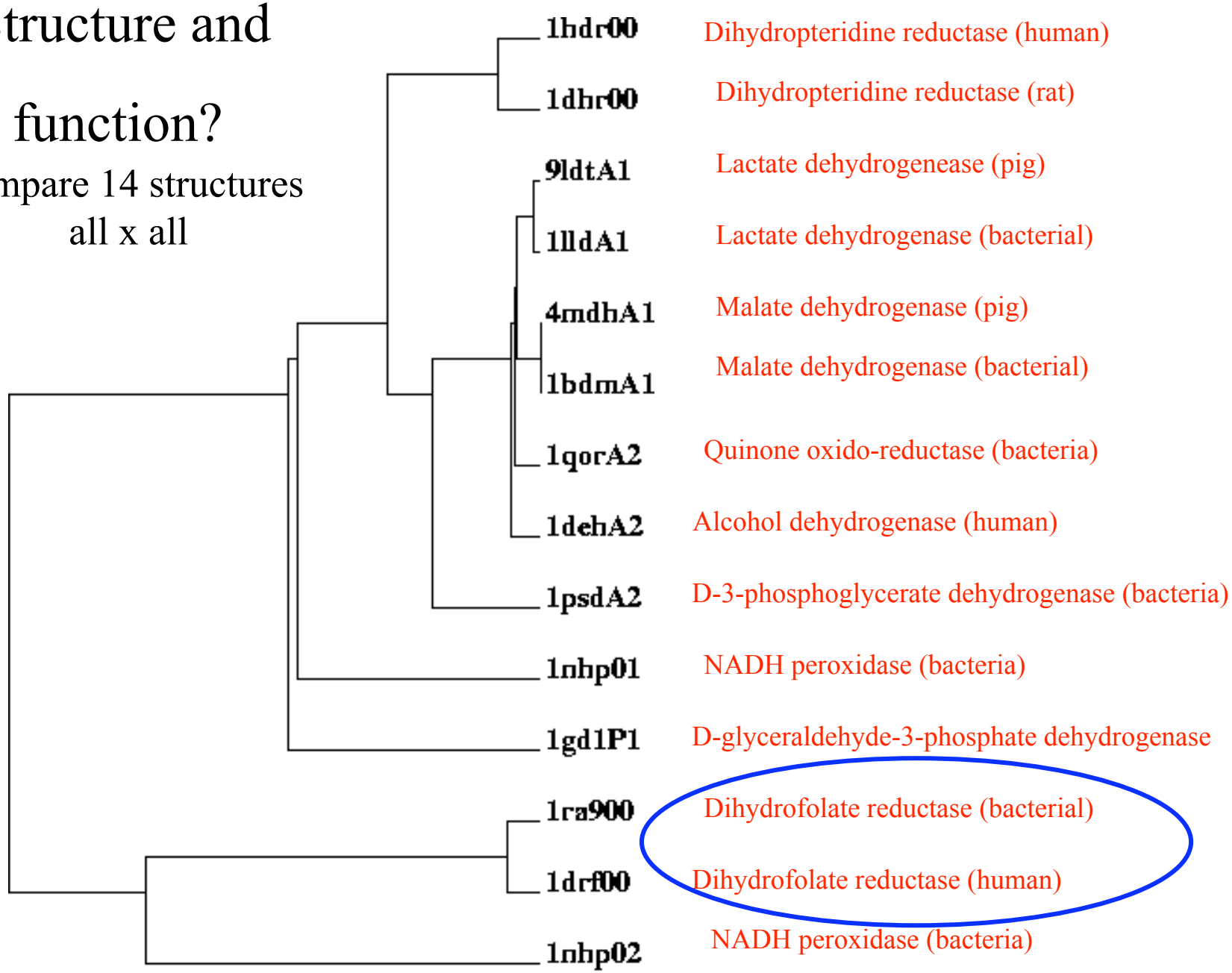
E.coli (1ra9)

Structure comparison and structural alignment

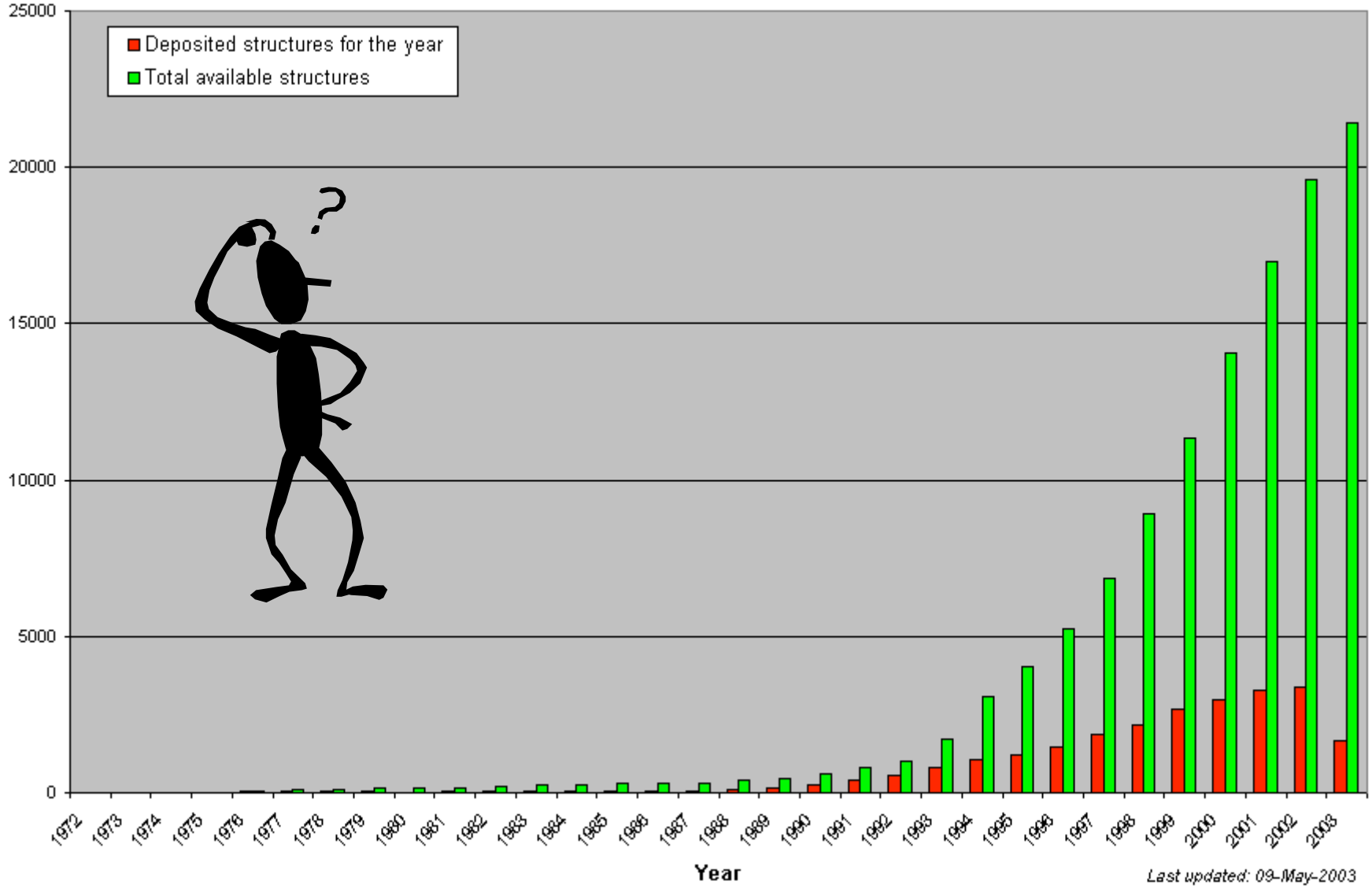
New methods from Computer Science…
Fast, accurate, biologically meaningful?
(…*prediction* of structure & function...)

Structure and function?

Compare 14 structures all x all

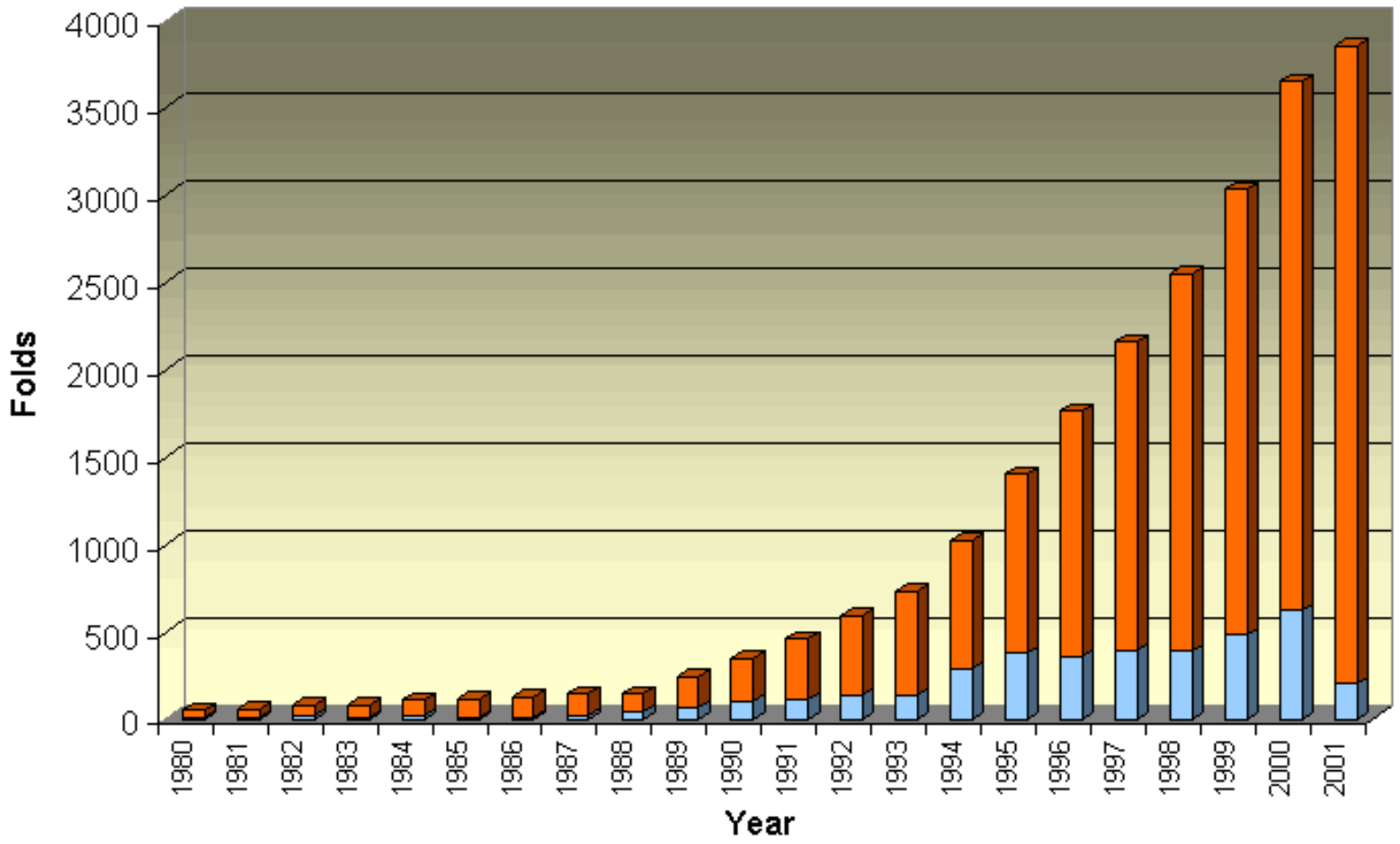


PDB: Protein Data Bank



www.rcsb.org/pdb/holdings.html

PDB Growth in New Folds



www.rcsb.org/pdb/holdings.html

PDB Holdings

		Molecule Type				
		Proteins, Peptides, and Viruses	**Protein / Nucleic Acid Complexes**	**Nucleic Acids**	**Carbohydrates**	**total**
Exp. Tech.	**X-ray**	16339	782	670	14	17805
	NMR	2590	90	518	4	3202
	Total	18929	872	1188	18	21007

SCOP: 1.61 (Sep 2002): 17406 PDB entries, 44327 domains

CATH 2.4 (Jan 2002): 36480 domains

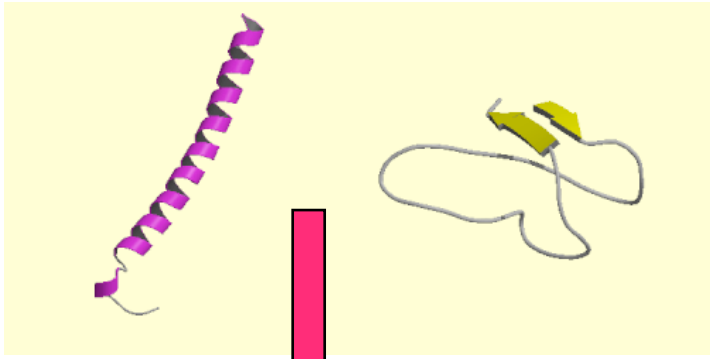
All x all comparison (SCOP): 10^9 comparisons @ 1 sec/comparison ~= 30 years...

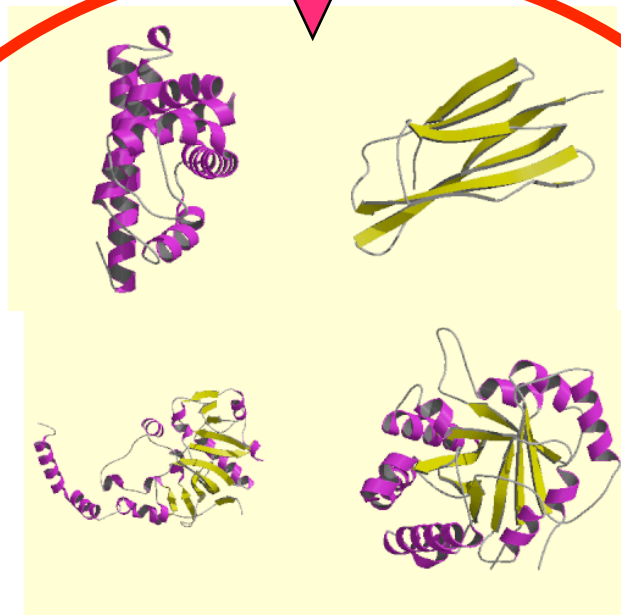
Protein structure - levels

PRIMARY STRUCTURE (amino acid sequence)

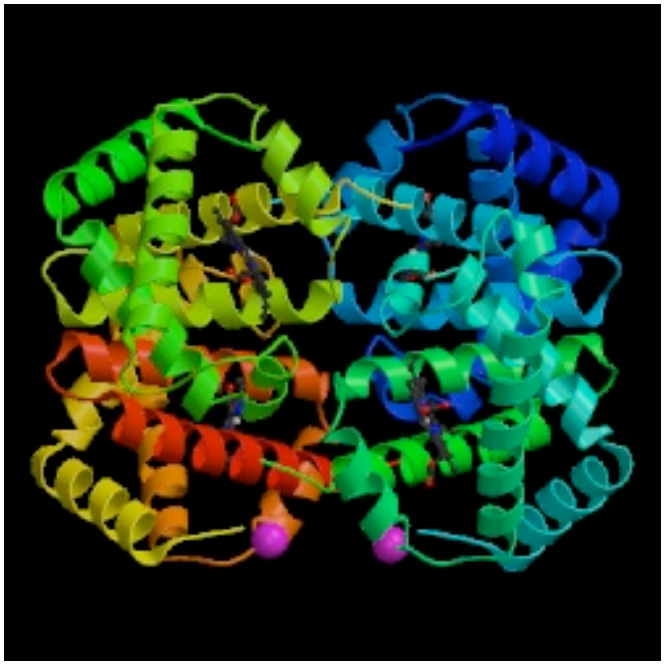
VHLTPEEKSAVTALWGKVNVDE
VGGEALGRLLVVYPWTQRFFE
SFGDLSTPDAVMGNPKVKAHG
KKVLGAFSDGLAHLDNLKGTFA
TLSELHCDKLHVDPENFRLLGN
VLVCVLAHHFGKEFTPPVQAAY
QKVVAGVANALAHKYH

SECONDARY STRUCTURE (helices, strands)



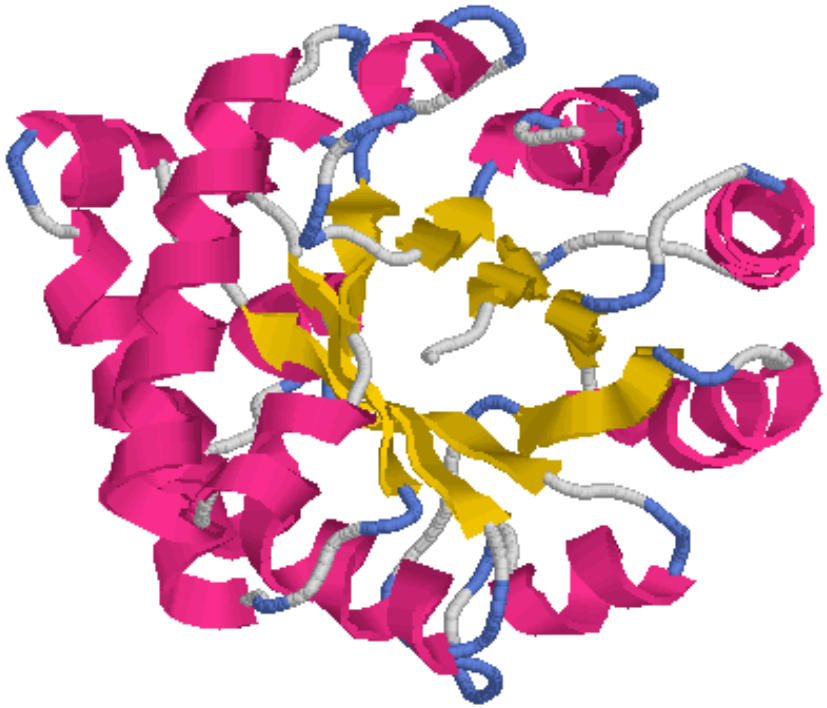


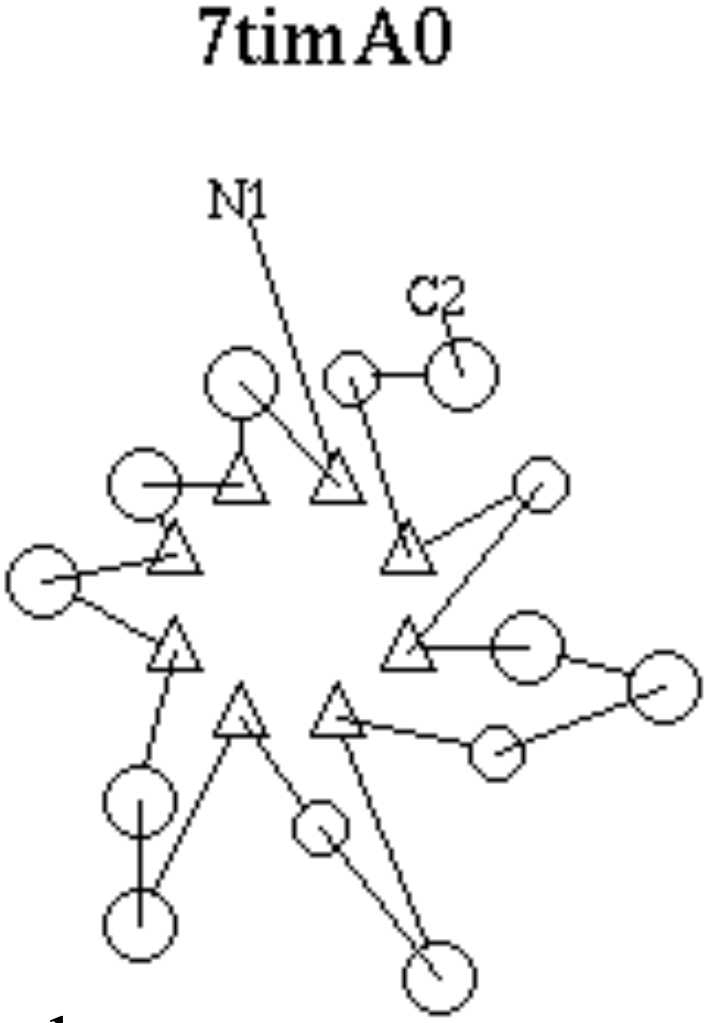
TERTIARY STRUCTURE (fold)

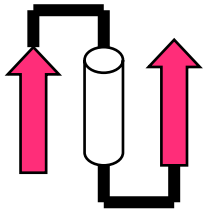


QUATERNARY STRUCTURE

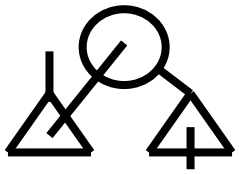
Protein structure







TIM barrel



TOPS

- Modelling & rapid analysis of protein structures
- Fold level: high level description of protein structures
  - Discover common motifs for protein families.
  - Very fast protein structure comparison system
- UK joint funded project: Glasgow & Leeds
- Enhanced database for topological descriptions of protein structures,
  - amino-acid sequence
  - several types of long-range interactions
  - ligand binding sites.

Latest developments

- Developing integrative machine learning methods to discover powerful motifs ("knowledge patterns")

- Relate the topological description of structures to
  - characteristics of the sequence,
  - functional characteristics of the protein (...ligand binding sites).

- Ultimate aim: prediction of protein structure and function.

TOPS

Simplified descriptions of protein 3D structures and their use in searching and structural pattern recognition

- TOPS diagrams ("cartoons") - optimal projection of protein 3-dimensional structures in 2 dimensions.
- Originally used for understanding & manual comparison of protein folds.
- Originally drawn manually [Sternberg+77]
- Now produced automatically from PDB files [Flores94, Westhead98]
- On-line atlas, search by protein domain name.
- Computation using TOPS descriptions

What can we do with TOPS? (just pretty pictures...)

- Structure comparison
- Detection of common structural motifs

$\Rightarrow$ Requires

- motif *discovery*
- motif *matching*

"Motif" = ***pattern*** with biological meaning

Eidhammer, Jonassen & Taylor, "Structure Comparison and Structure Patterns", JCB, 7:5 685-716, 2000.

Topological description

- Consider sequence of SSEs (strand, helices), plus spatial adjacency within fold & approximate orientation
- Neglect details (lengths & structures of loops, exact lengths & spatial orientations of SSEs, ...)

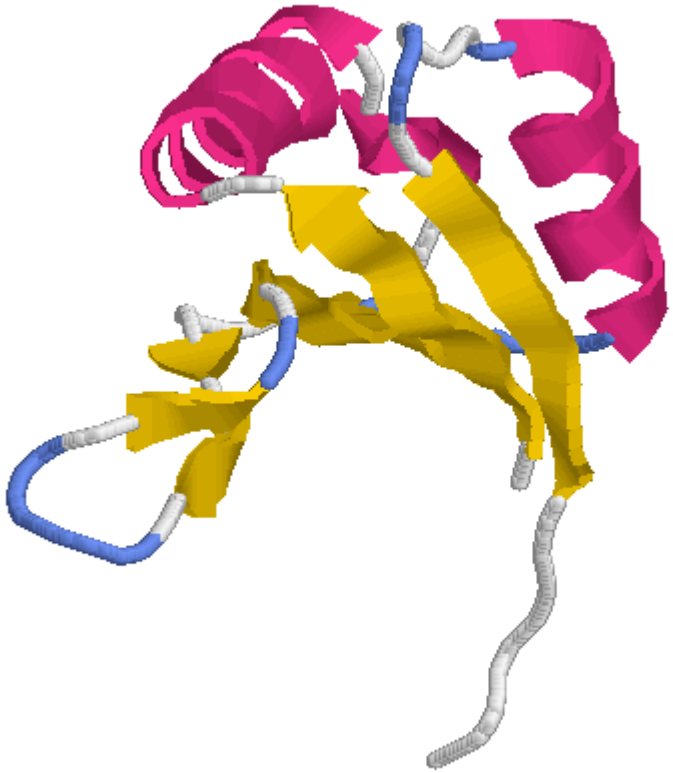
√ simplicity

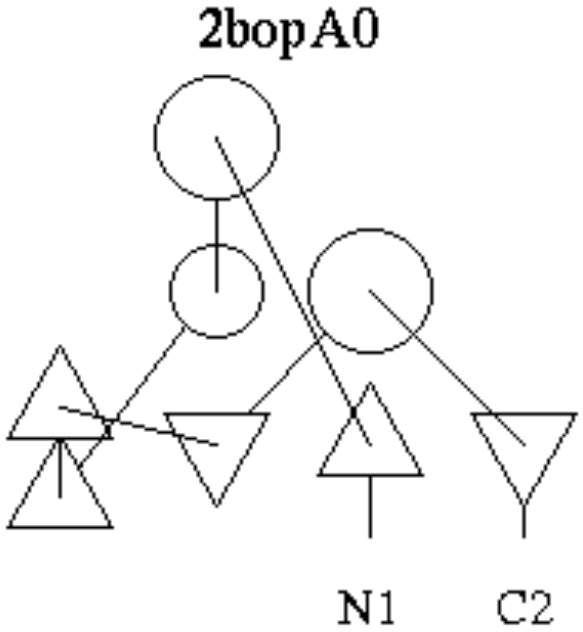
  - implement very fast comparison algorithms, machine learning, ...
  - detect distant structural relationships

X simplicity

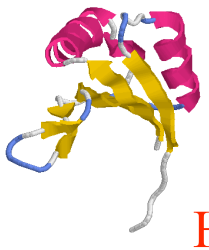
  - relate structures topologically which have no meaningful biological relationship.

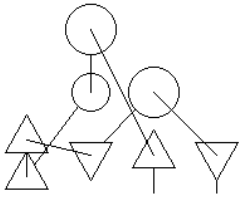
Example - a plait

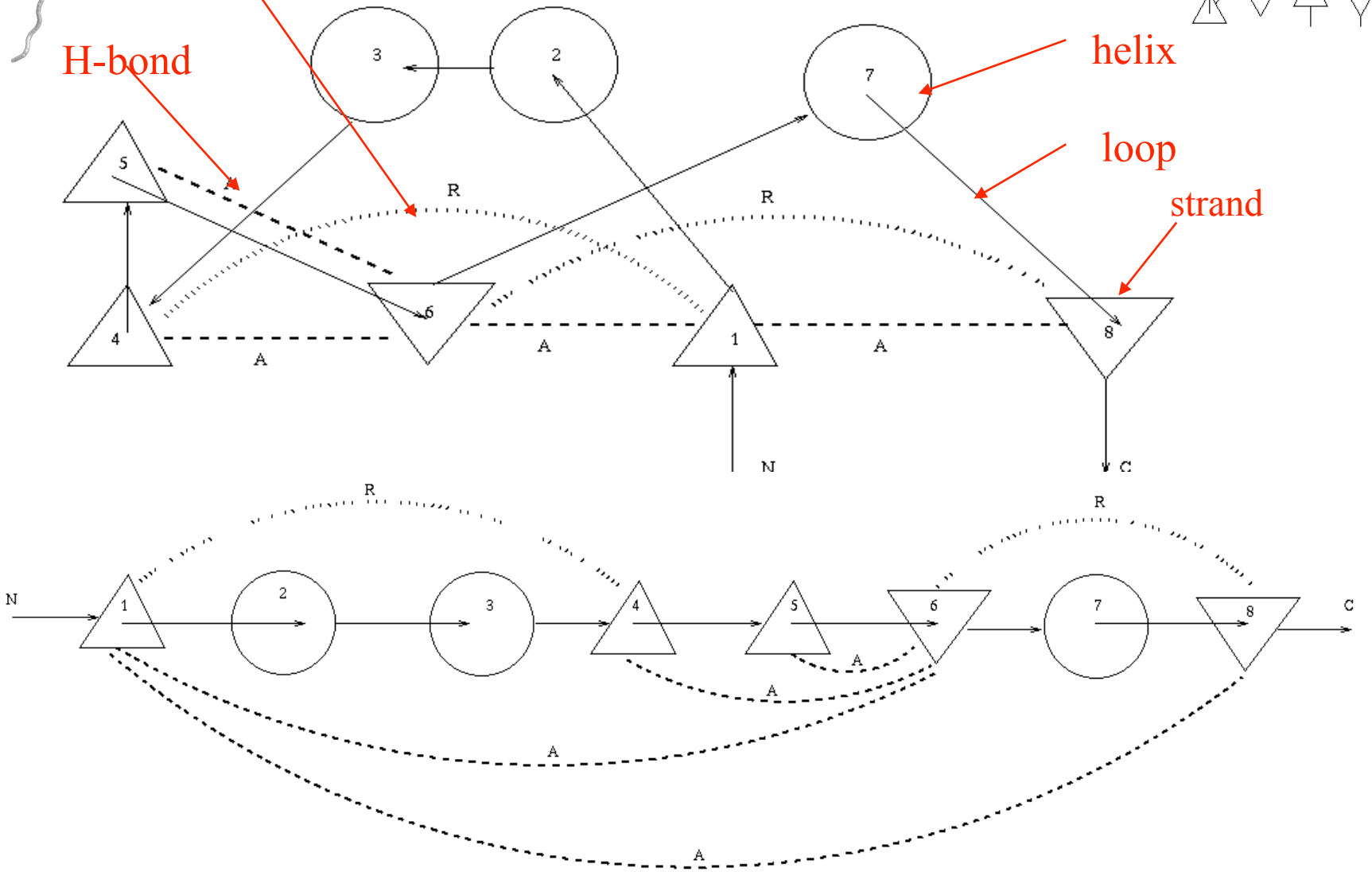




2bopA0








```
2bop NEhhEEeHeC 1:4R 1:6A 1:8A 4:6A 5:6A 6:8R
```

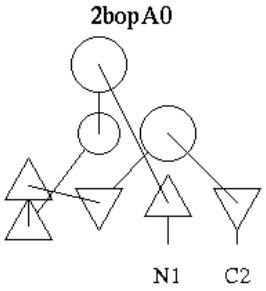
Generating TOPS descritpions

PDB file

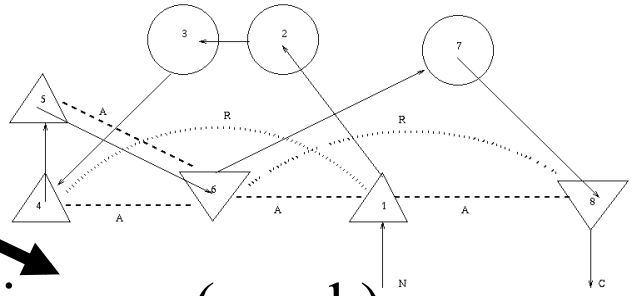
DSSP file

Automatic layout

Edit layout ⇄ TOPS file



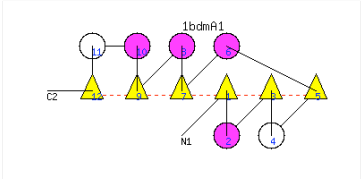
TOPS cartoon
(picture)

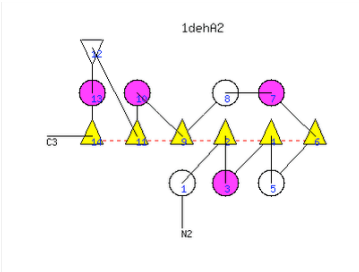


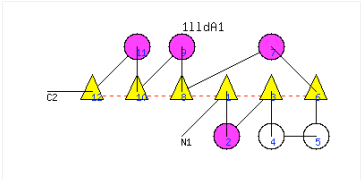
TOPS diagram (graph)

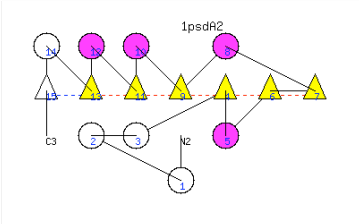
What is a pattern?

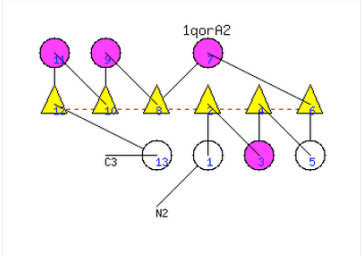
- Several examples, with common parts highlighted

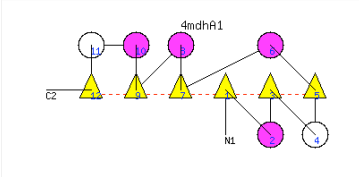


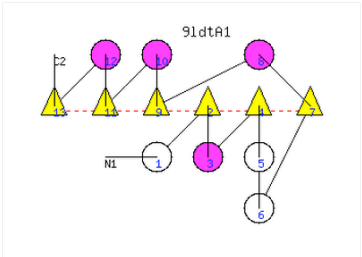




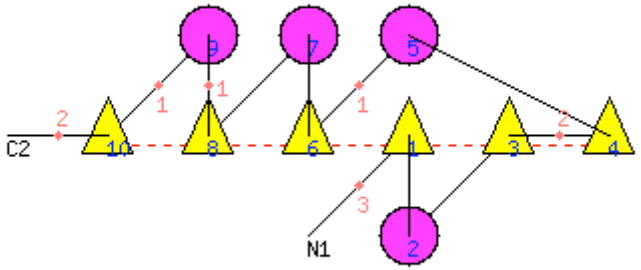




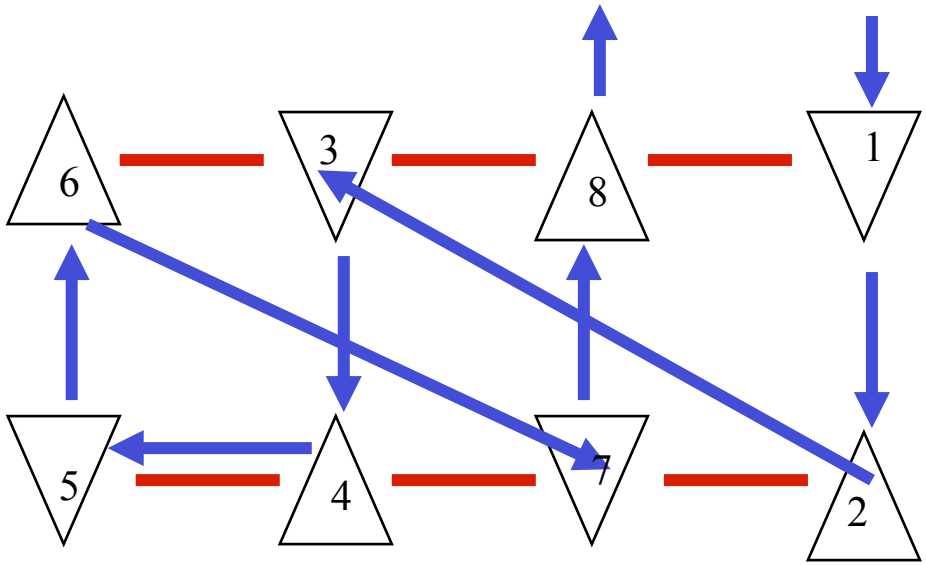




- A common description

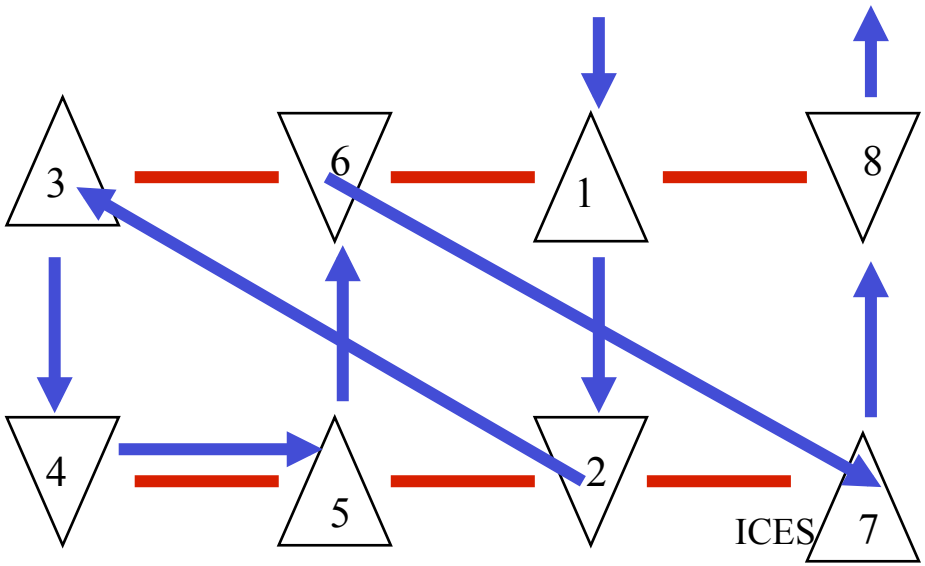


“Jelly roll” motifs (anti-parallel β-sandwich)



Jelly Roll 1

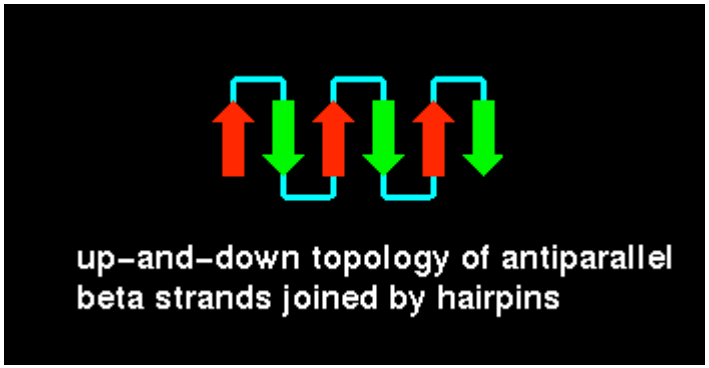
6-3-8-1
5-4-7-2

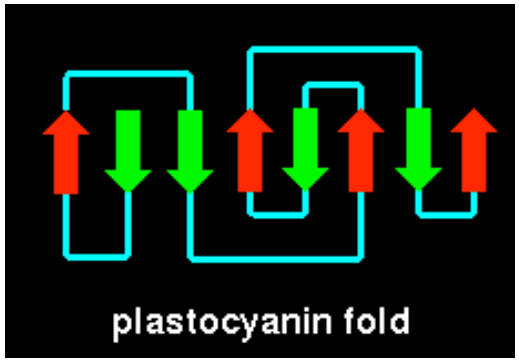


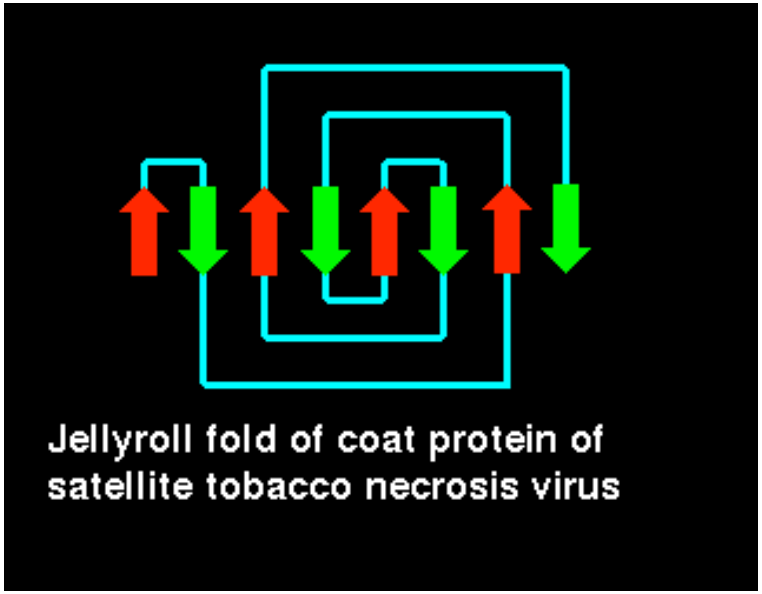
Jelly Roll 2

3-6-1-8
4-5-2-7

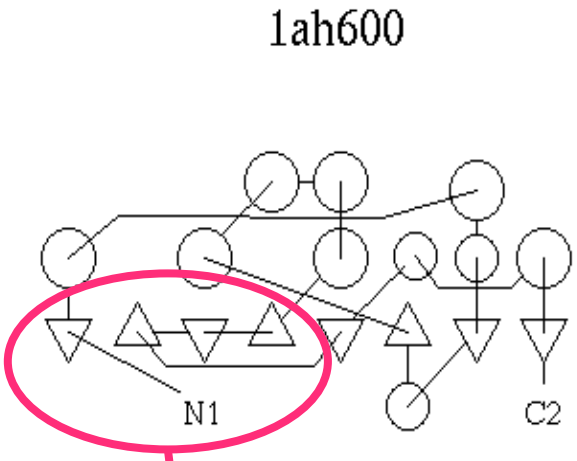
Some motifs: beta-sheet connectivities

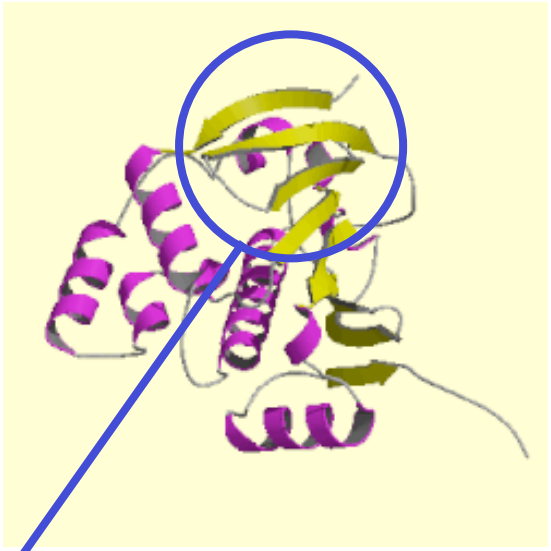


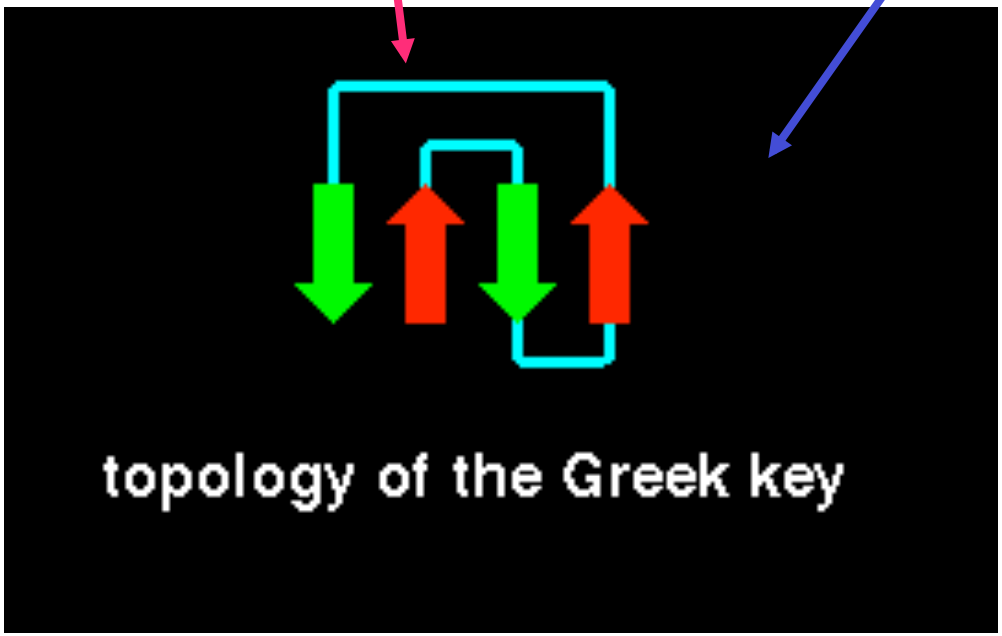




“Greek key” motif

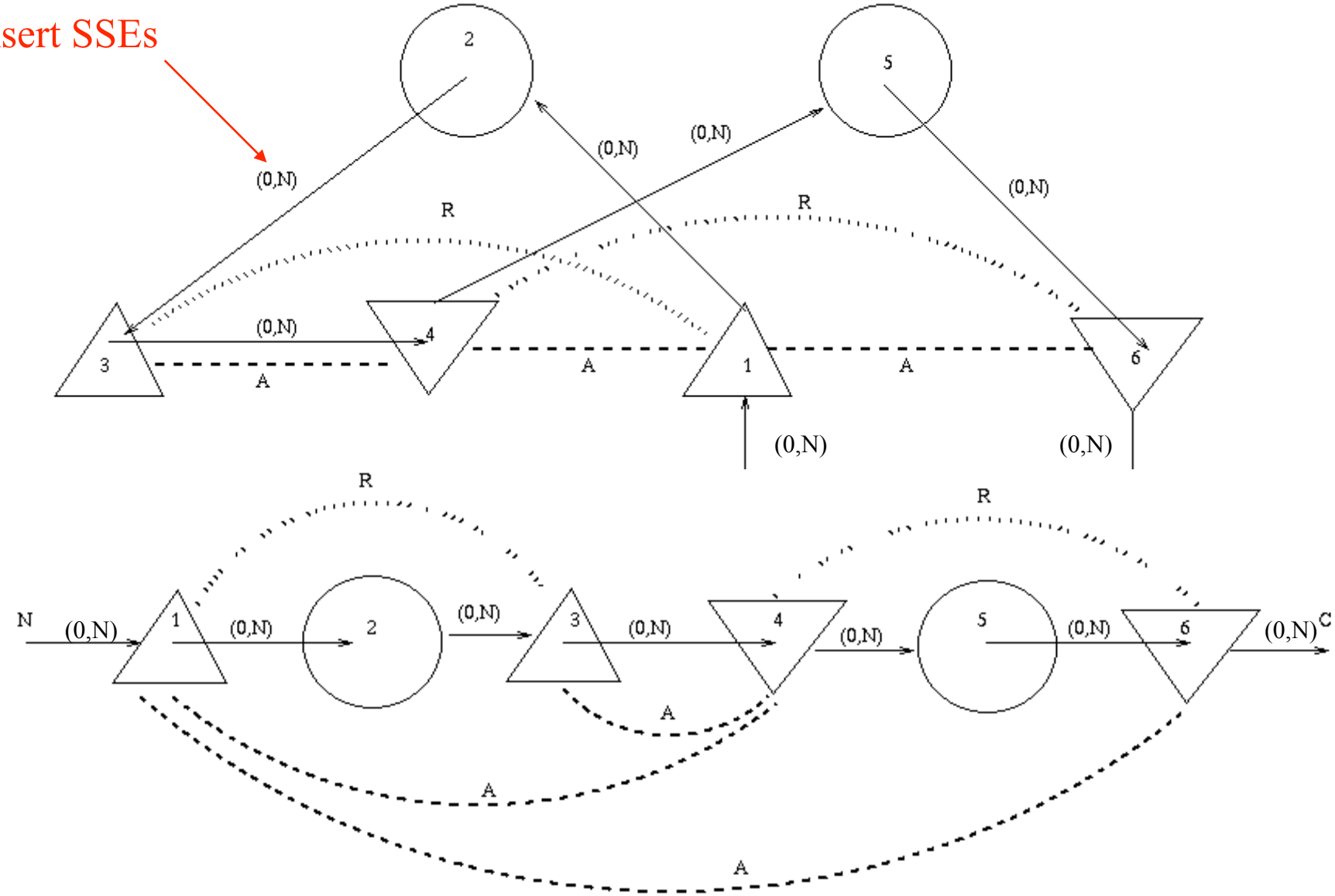






Plait motif

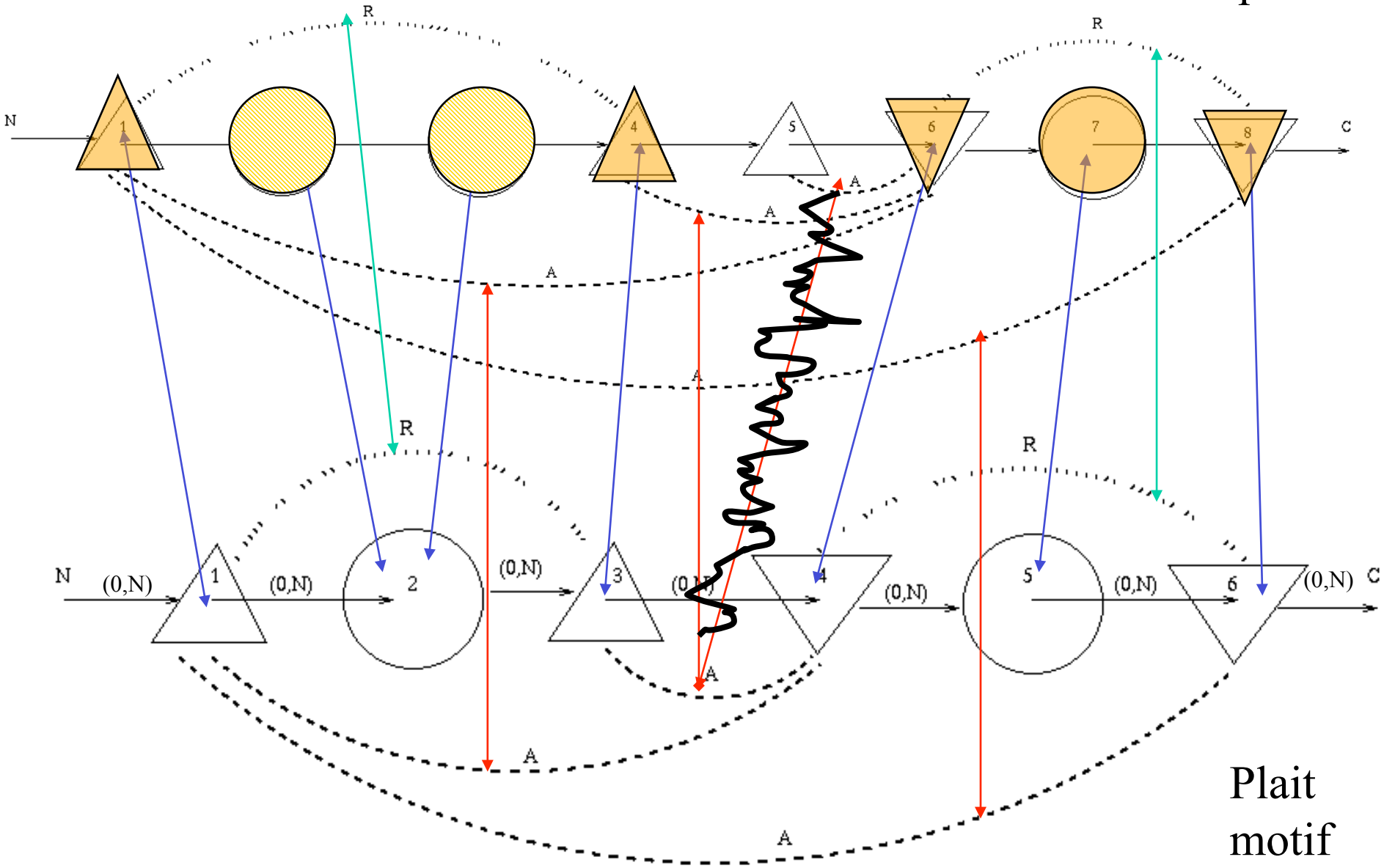
Number of insert SSEs




```
Plait E*h*E*e*H*e 1:3R 1:4A 1:6A 3:4A 4:6R
```

Pattern matching

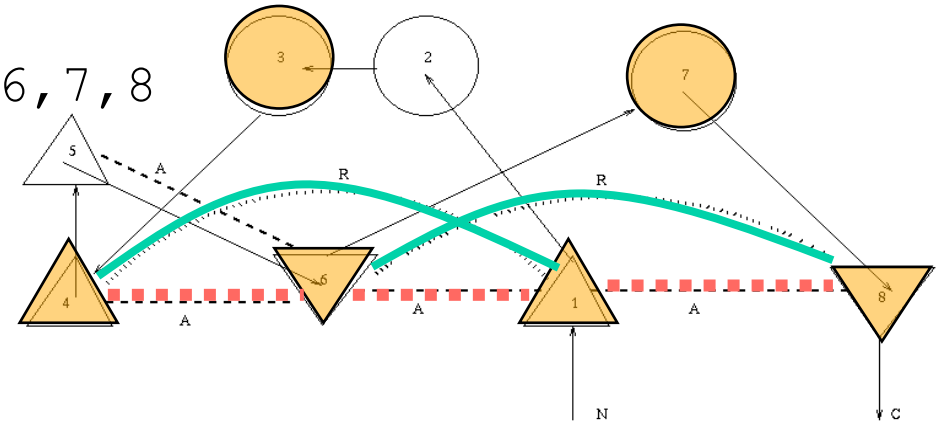
2bop



Plait motif

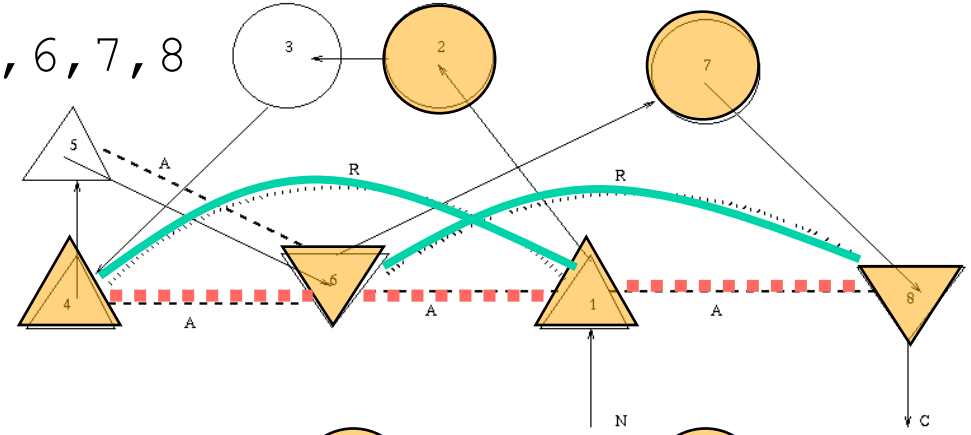
Alternative matches

1,3,4,6,7,8

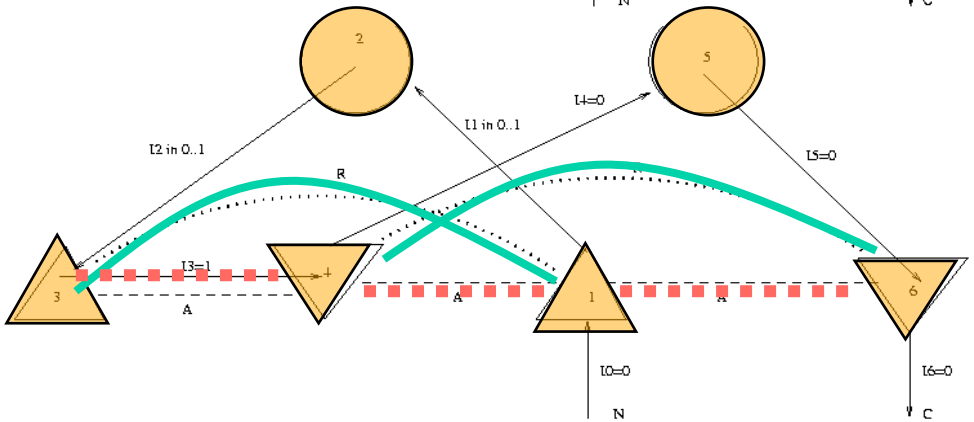


2bopA0

1,2,4,6,7,8



Plait



Topological search with ***plait motif***

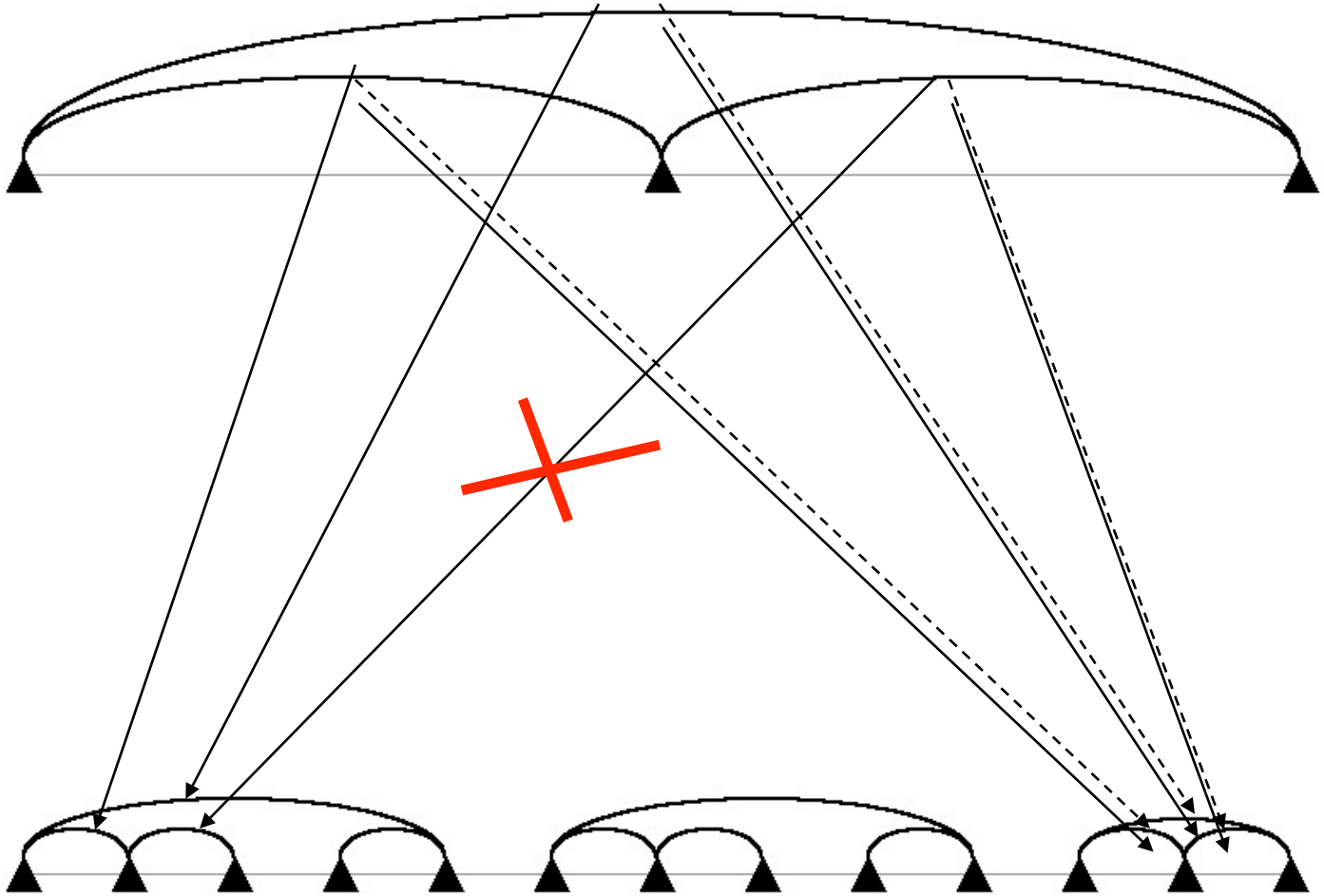
3567 match attempts, Search time: 4 sec, **1ms/match**
Total matched = 44

inserts	domain	CATH code
3	1pytA0	3.30. 70.340.1
3	1qba01	2.60. 40.290.2
4	1ab8A0	3.50. 6. 10.1
4	1ha101	3.30. 70.330.1
4	1ha102	3.30. 70.330.2
4	1numA0	3.30. 70.330.4
4	1urnA0	3.30. 70.330.4
4	2bopA0	3.30. 70.330.5
4	5rubA1	3.30. 70.150.1
5	1cg2A2	3.30. 70.360.1
5	1regX0	3.90.130. 10.1
5	1vaoA2	3.30.465. 20.1
5	1vhiA0	3.30. 70.390.1

...

New fast subgraph isomorphism algorithm

Pattern graph



Target graph

Pattern searches on TOPS databases

- Hand generated descriptions have written of well-known motifs:
  - left-handed βαβ
  - Greek keys
  - jelly rolls
  - plaits
  - Rossmann type NAD-binding domains
  - immuno-globulins
  - barrels and Tim-barrels (perfect and distorted)
  - trefoils
  - propellers
- User defined patterns
- Automatic pattern discovery

Approaches to pattern discovery

- **Pattern (language) driven:**

  enumerate all (or some) patterns up to certain complexity (length), for each calculate the score, and report the best

- **Comparison based:**

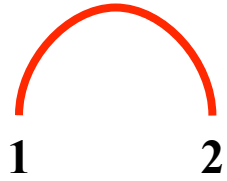
  find patterns by pairwise comparison of input objects

  A. Brazma, I. Jonassen, I. Eidhammer and D.R. Gilbert "Approaches to the automatic discovery of patterns in biosequences.", Journal of Computational Biology. Vol 5, Nr. 2, pp 277-303, 1998

Topological pattern discovery
(pattern extension and repeated matching)

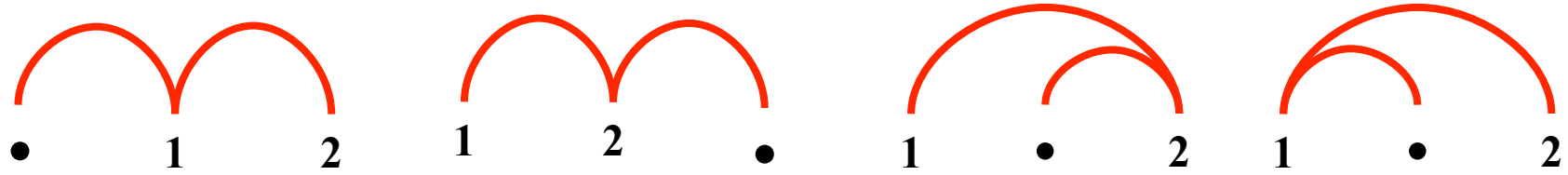
Repeat:

- Find new sheet

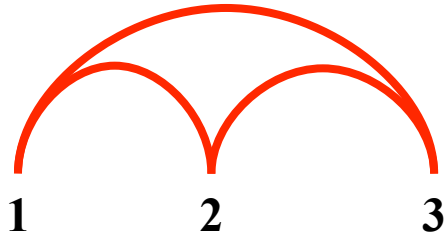


Works (in theory) on set of any size

- Extend current sheet

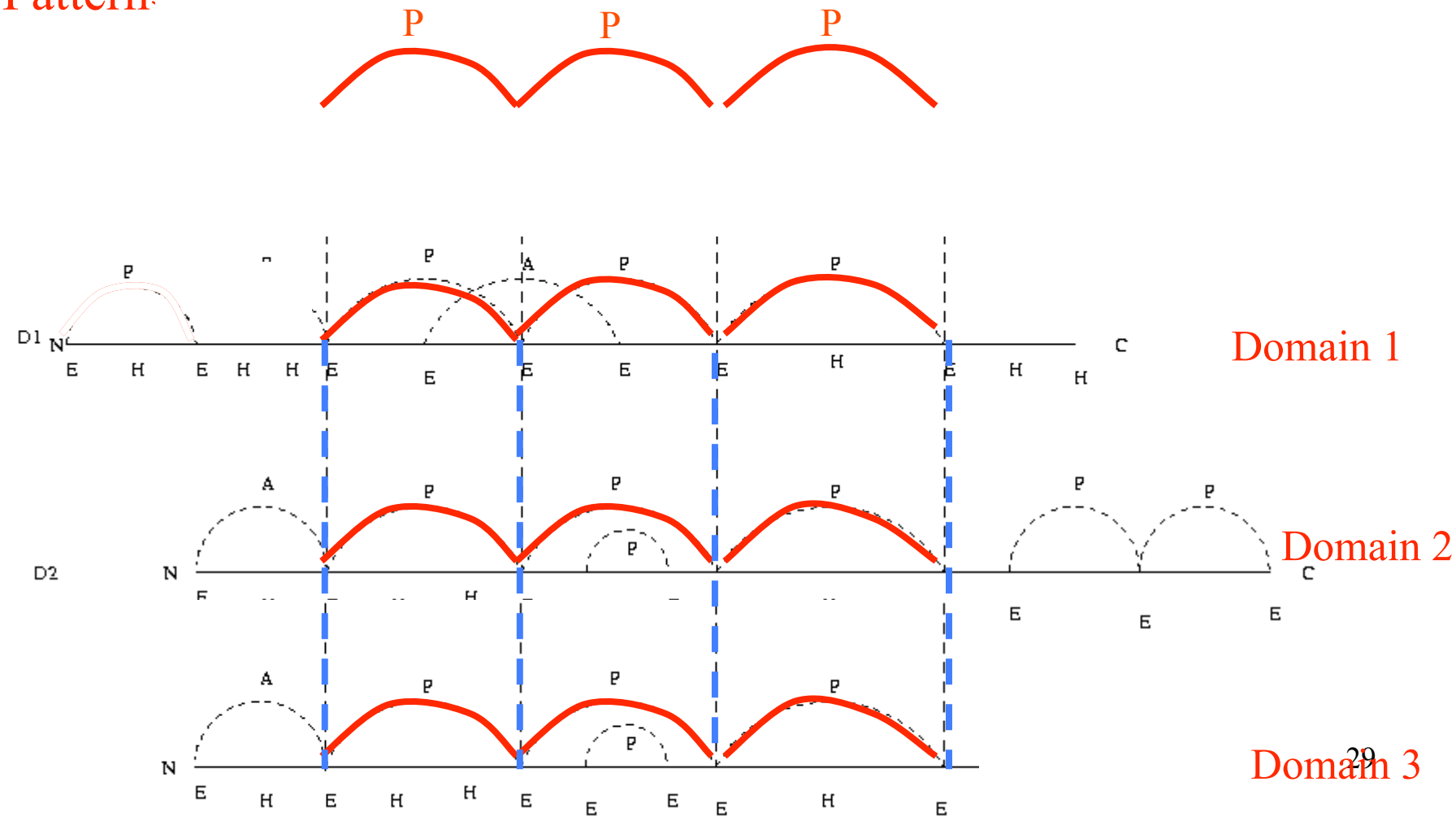


- Find circuits

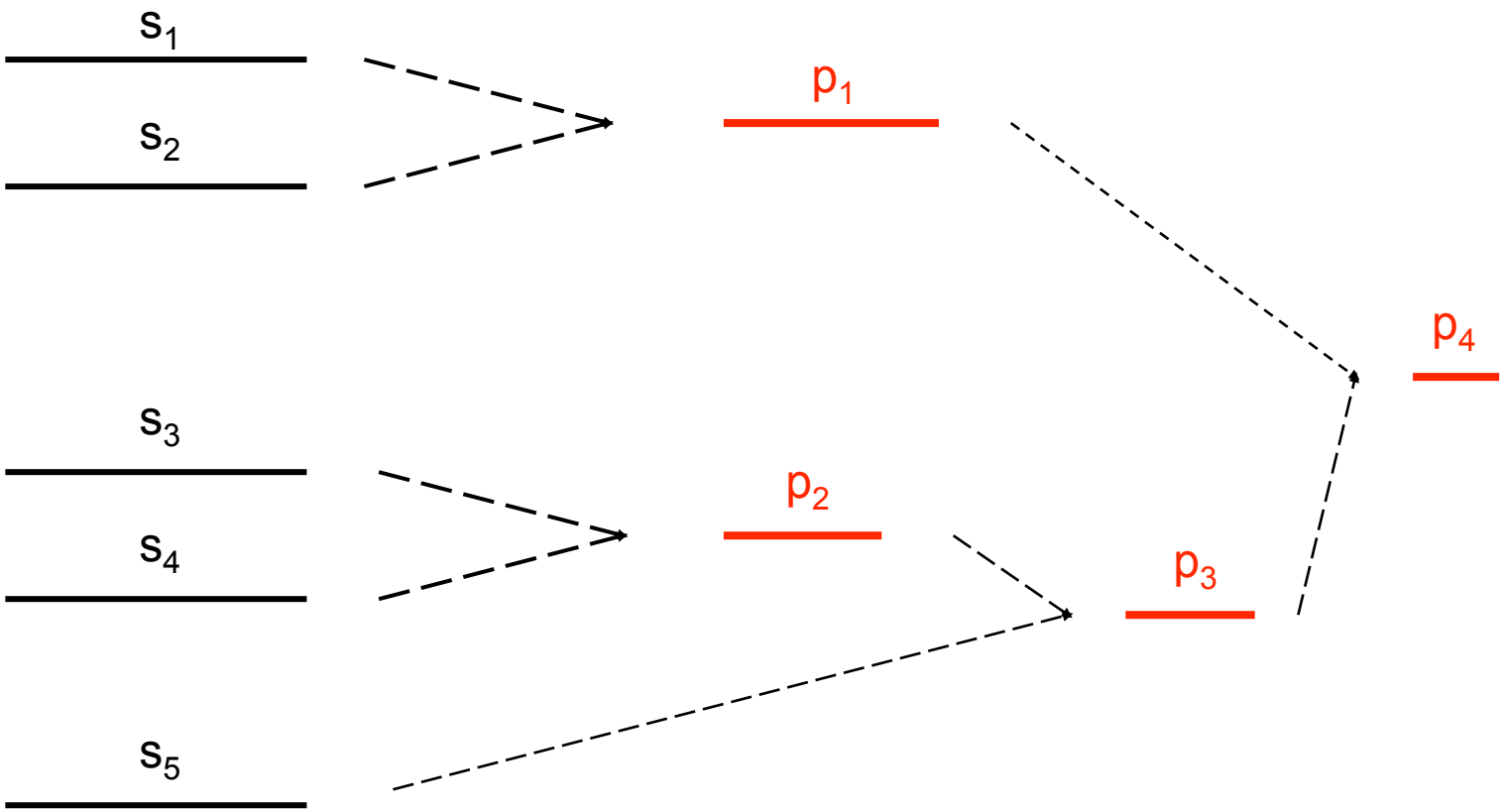


Discovering common patterns and making multiple alignments

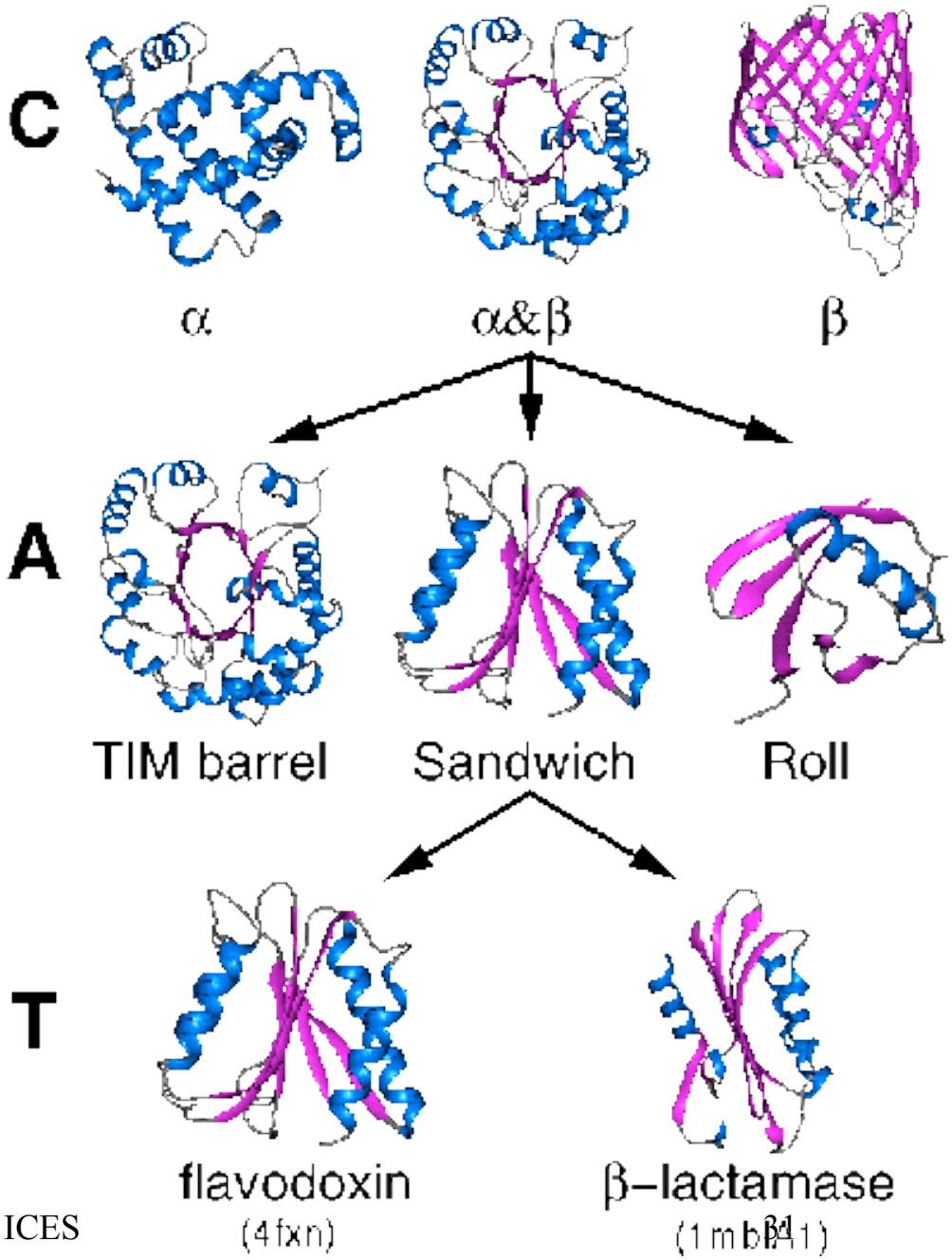
Pattern



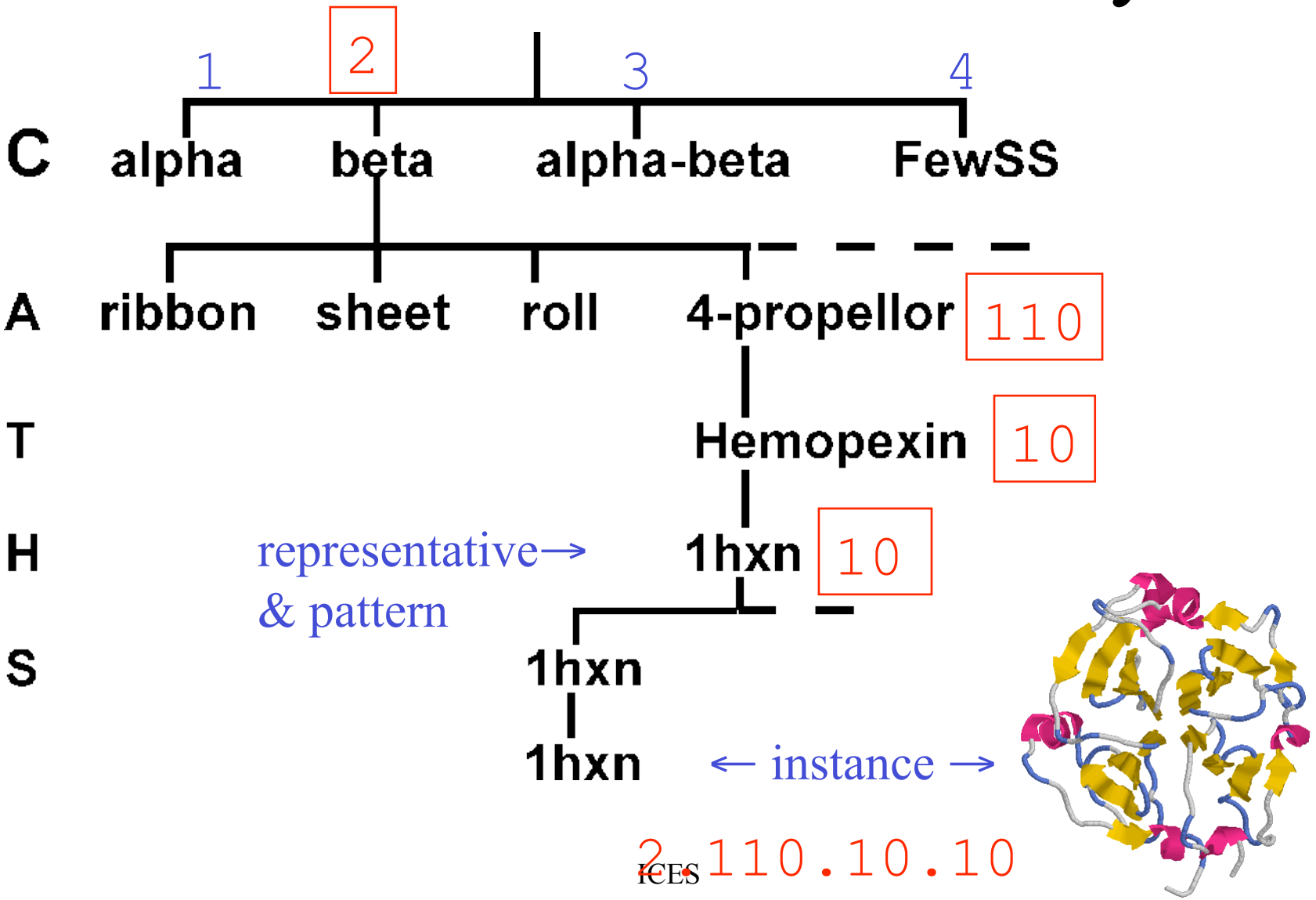
Comparison based approach



Back to patterns -
use in classification:
the CATH hierarchy

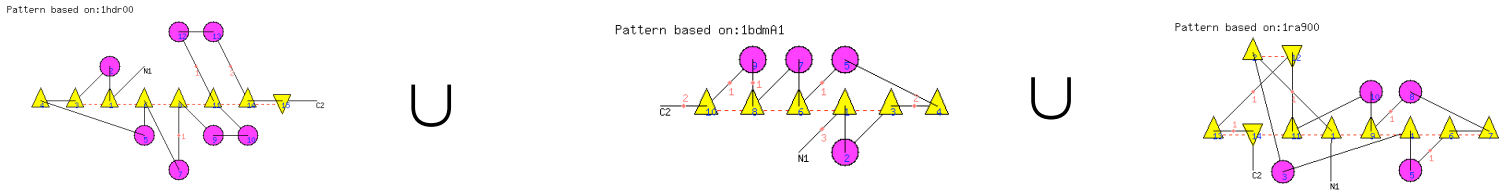


CATH database hierarchy



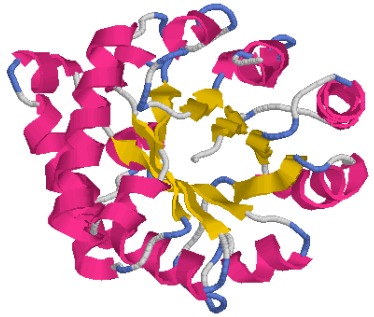
Extending TOPS patterns to unions

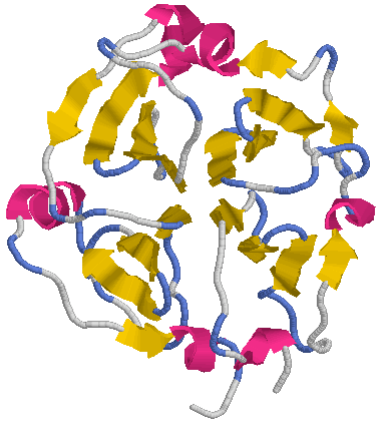
- Pattern = $pattern_1 \cup pattern_2 \cup \ldots \cup pattern_n$



Structure comparison

- Atomic coordinate level (RMSD)
- Threading and double dynamic programming
- Graph comparison
- Alignment using discovered patterns
- Issues:
  - validation (Gold Standard = Alexei Murzin)
  - distance metric (triangle inequality)





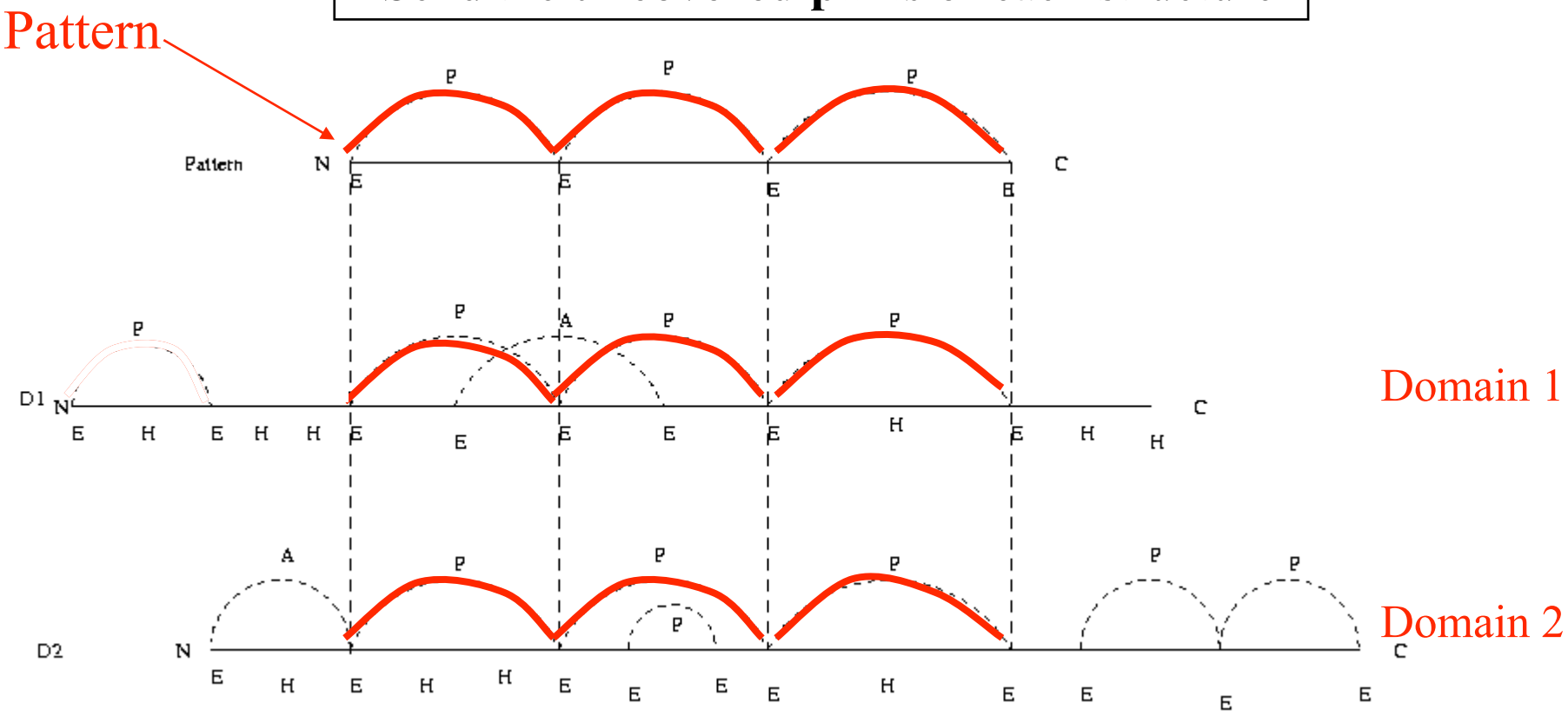
Comparison:
alignment using discovered patterns

Rating patterns

- Sensitivity, Specificity etc
- Size
  (e.g. number of SSEs, arcs etc)
- Compression
  (measure of how much of each of the items in the learning set is described)

Compression

- Send the pattern once, and
- Send the **uncovered parts** of each structure



Special case: When 2 examples, compression gives *comparison*

Structure comparison: 1 x all

User = xyz@..... Submitted at 19:55:53 on 31/03/03, Submitted file name = TEST Database = atlas

==========

Motifs found (SSE numbers):
Barrel with 8 strands, SSE numbers = [6,9,12,15,19,22,24,26] , 8 parallel, and 0 anti-parallel
Immuno-globulin with two anti-parallel sheets [32,31,5], [2,29,27]
Barrel : [9,12,15,19,22,24,26,6], Curved sheet: [35,34,5,31,32,33,36,3,4], Sheet: [27,29,2,1]

==========

3567 comparisons, Comparison time : 674 sec

Domain	Code	Rank
CODE	target_query	0
1ak500	3.20.20.320.1	21
2amg00	3.20.20.70.10	21
1gox00	3.20.20.220.1	22
1fcbA2	3.20.20.220.1	24
1ads00	3.20.20.130.1	29
1frb00	3.20.20.130.1	29
1ah400	3.20.20.130.1	30
1aj000	unknown	31
1pii02	3.20.20.30.1	31
2kauC2	unknown	31
1qbd00	unknown	32
1ubsA0	3.20.20.30.4	32
1dhpA0	3.20.20.60.3	34
1htiA0	3.20.20.80.1	34
1igs00	3.20.20.30.3	34
1nal10	3.20.20.60.1	34
1reqB1	unknown	34
1aw2A0	3.20.20.80.1	35

3.20.20.20

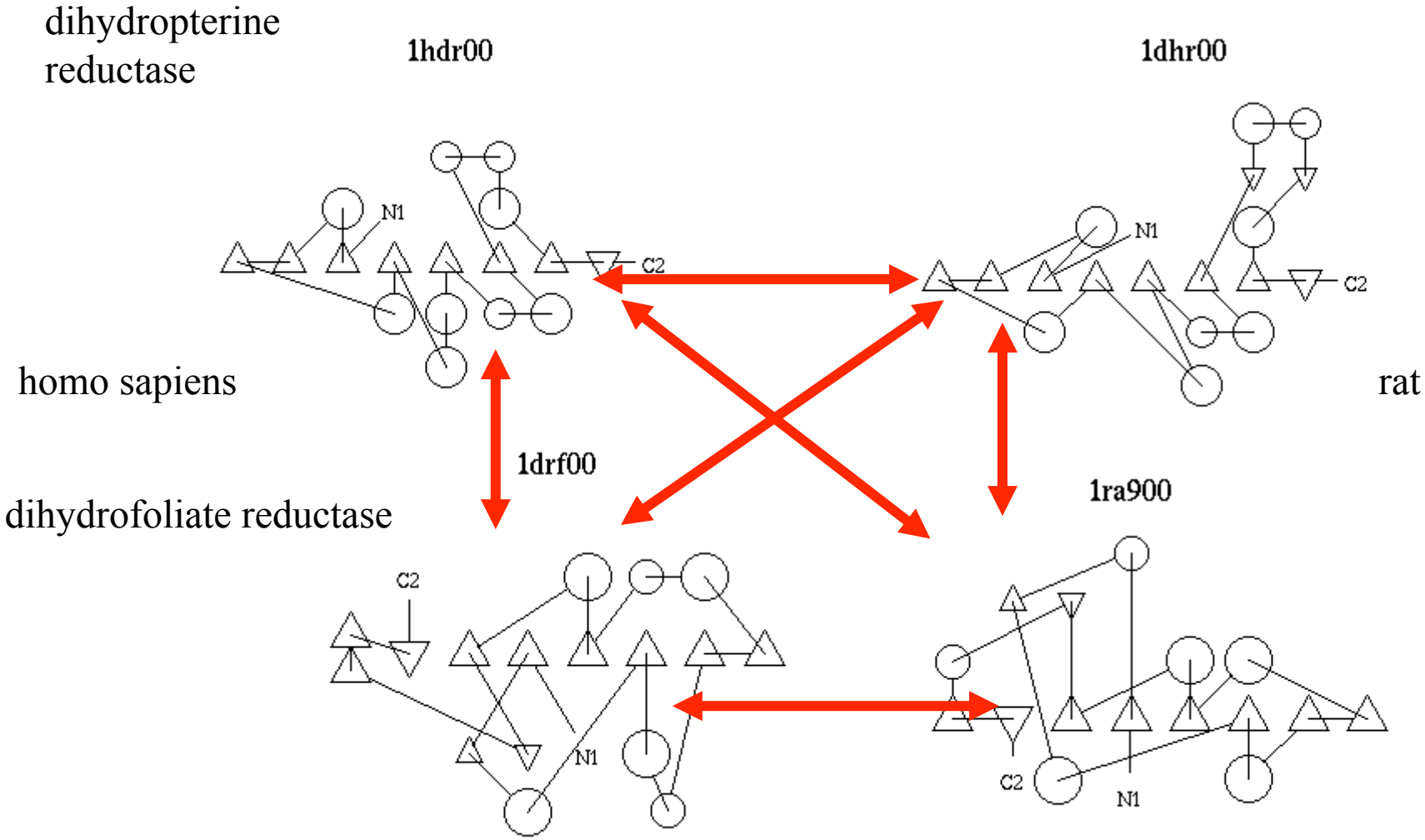
3.20.20.140

3.20.20.240

Structure comparison server

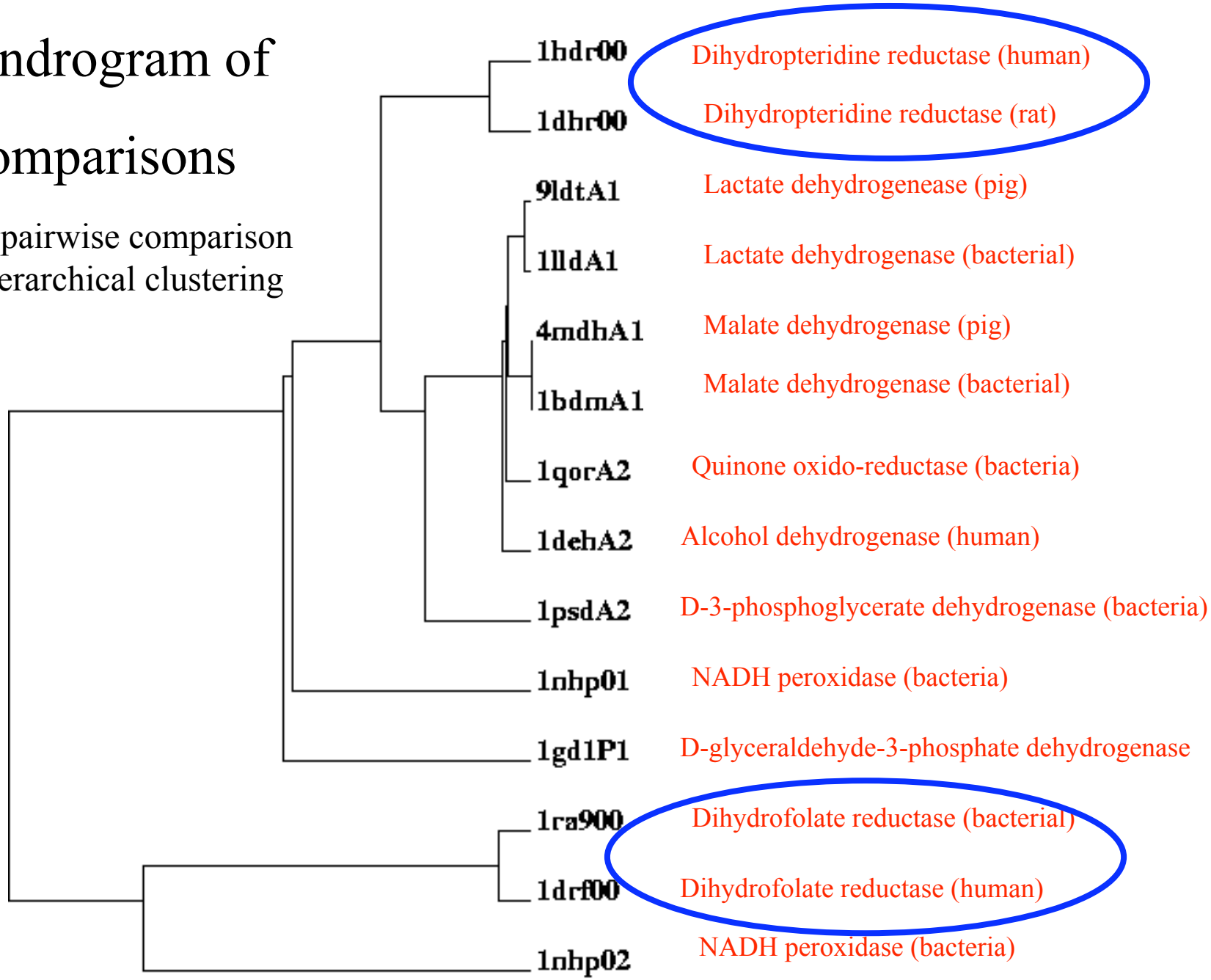
www.tops.leeds.ac.uk

Pairwise all x all comparison (NAD binding domains)

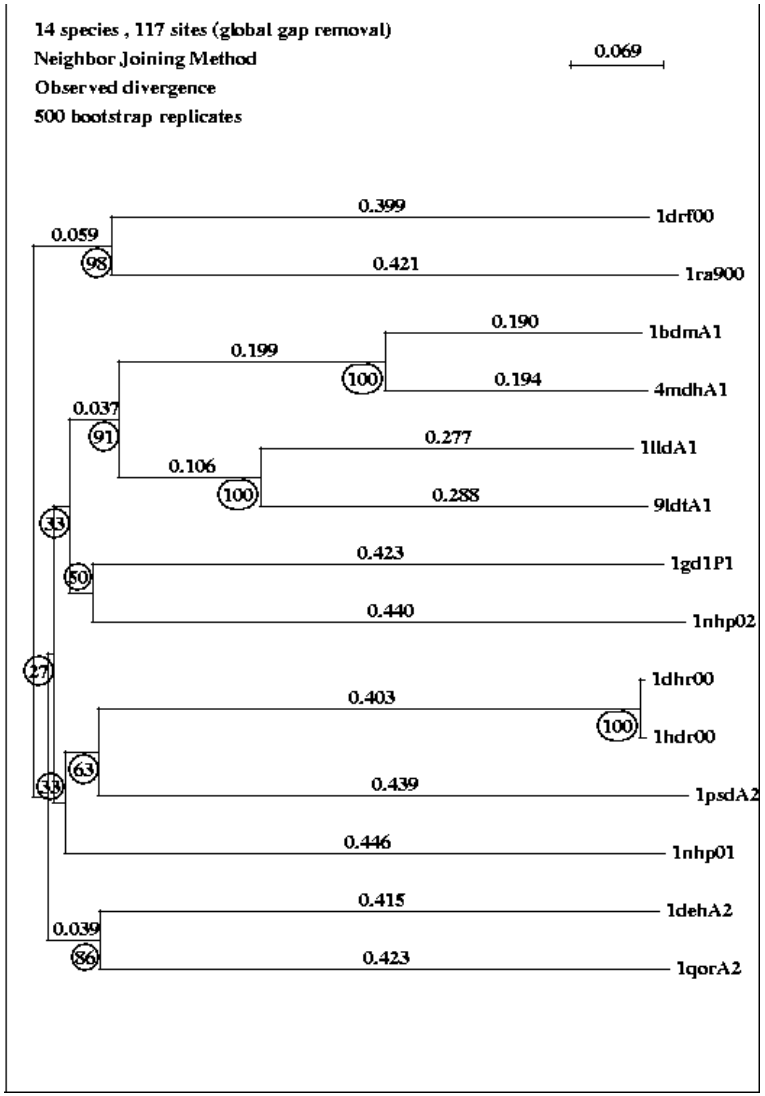


Dendrogram of comparisons

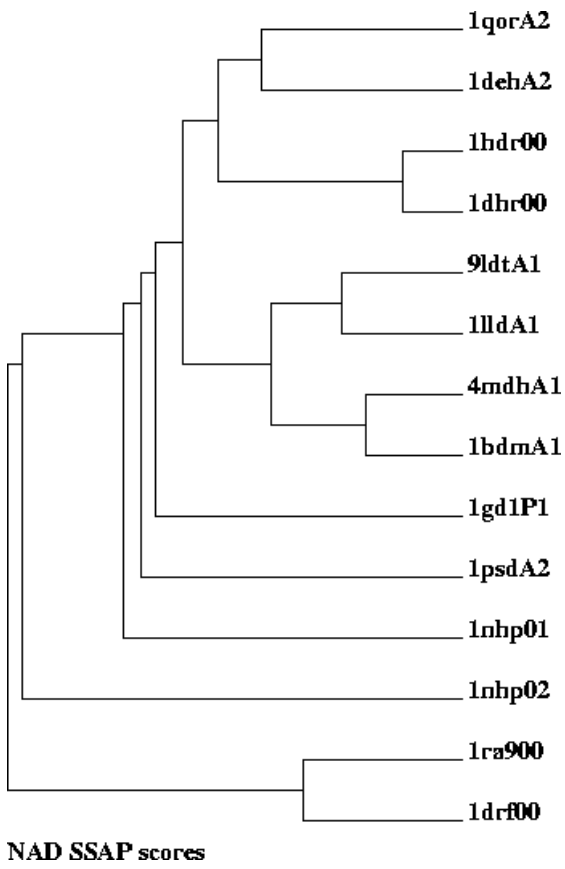
From pairwise comparison & hierarchical clustering



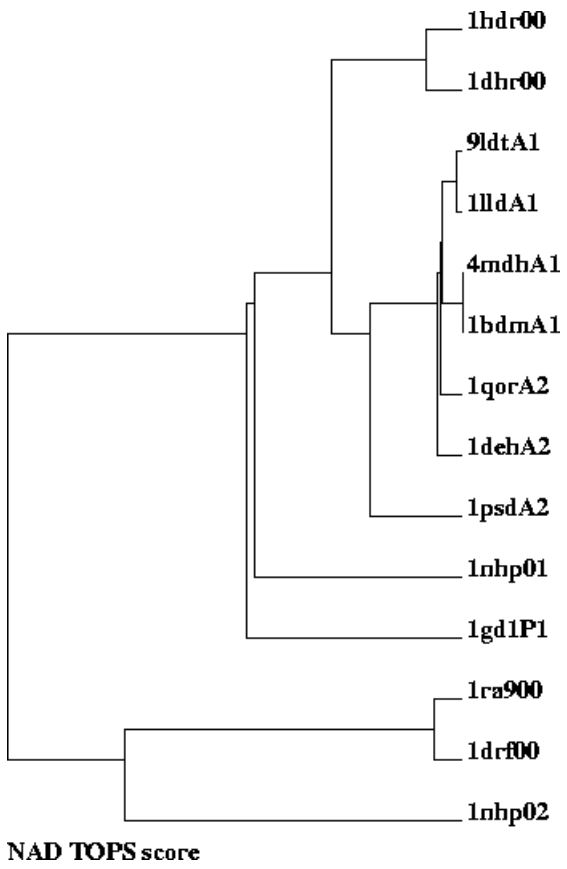
NAD comparisons



Sequence

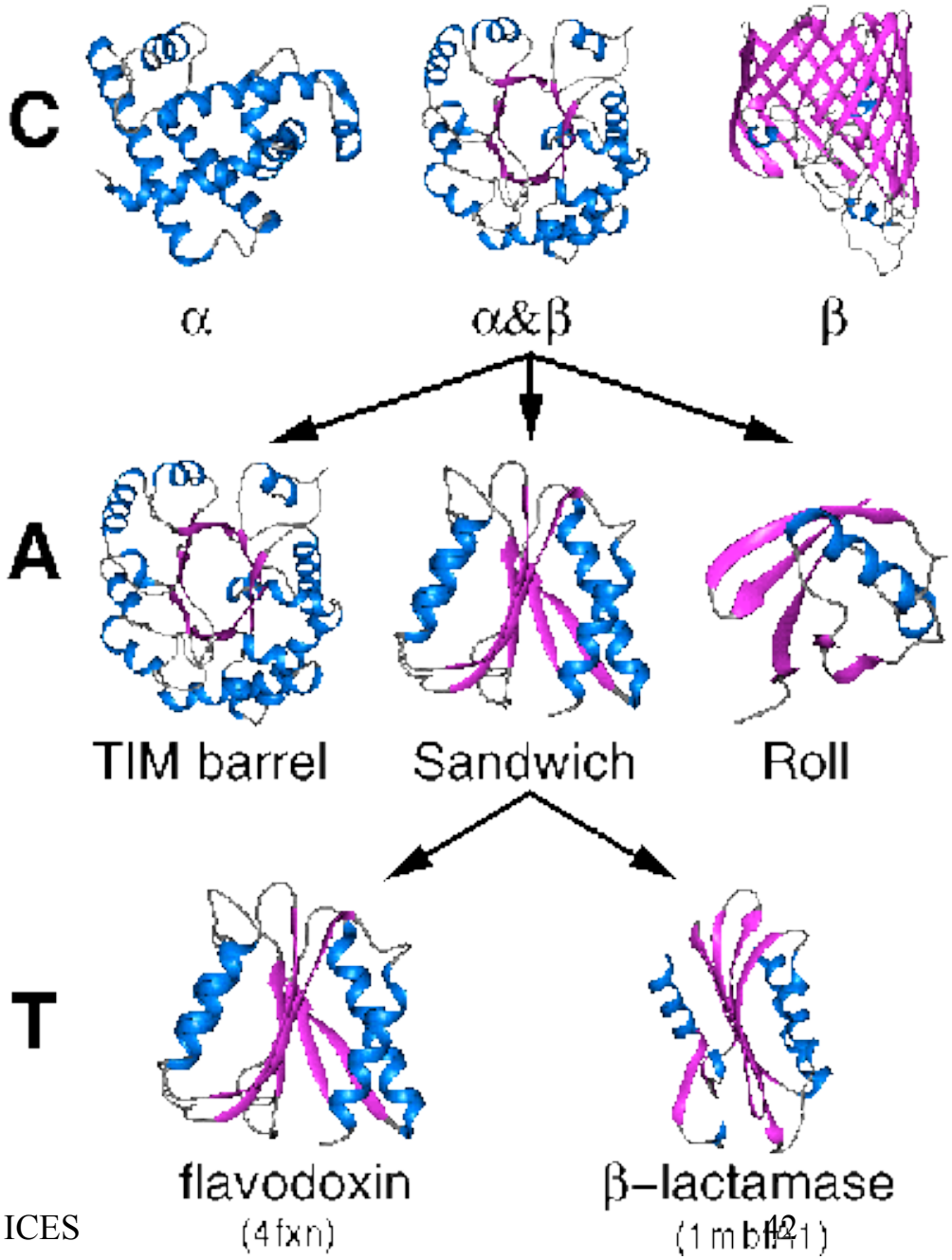


Structure
(atomic coordinates)

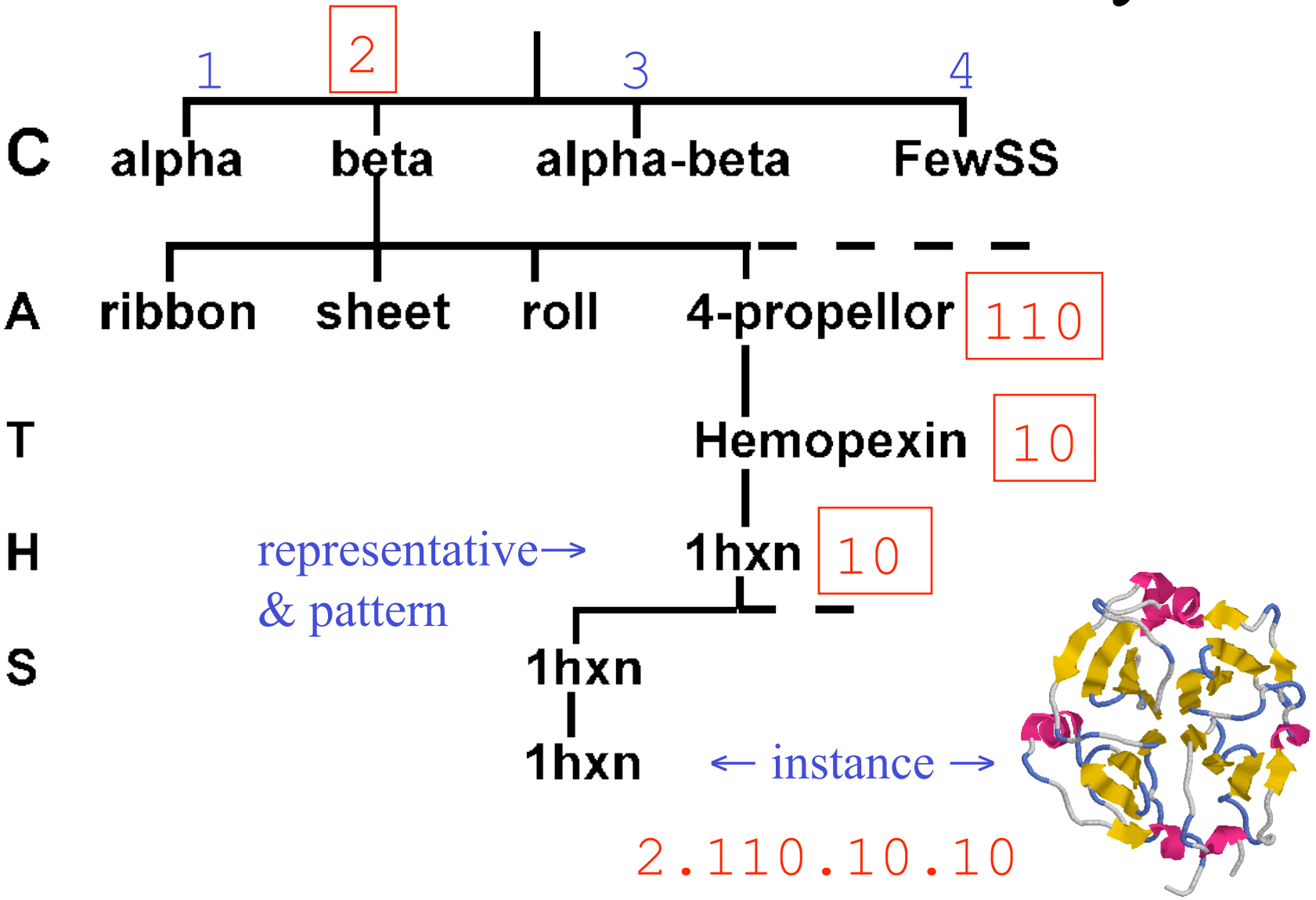


Structure
(topology)

Patterns - use in classification: the CATH hierarchy

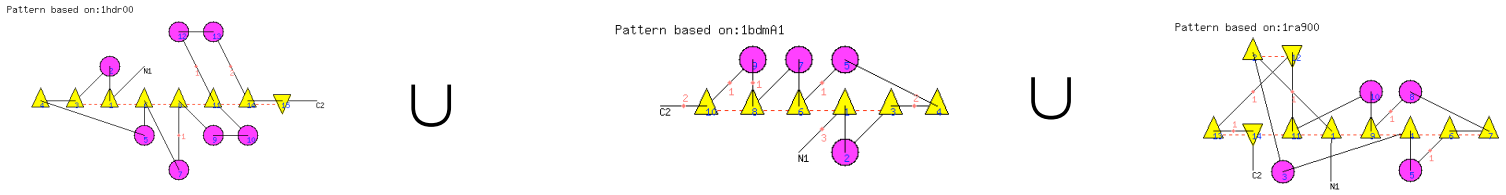


CATH database hierarchy



Extending TOPS patterns to unions

- Pattern = $pattern_1 \cup pattern_2 \cup \ldots \cup pattern_n$



Case study

GHKL an emergent ATPase/kinase superfamily

Dutta R, Inouye M, Trends Biochem Sci 2000 Jan;25(1):24-8

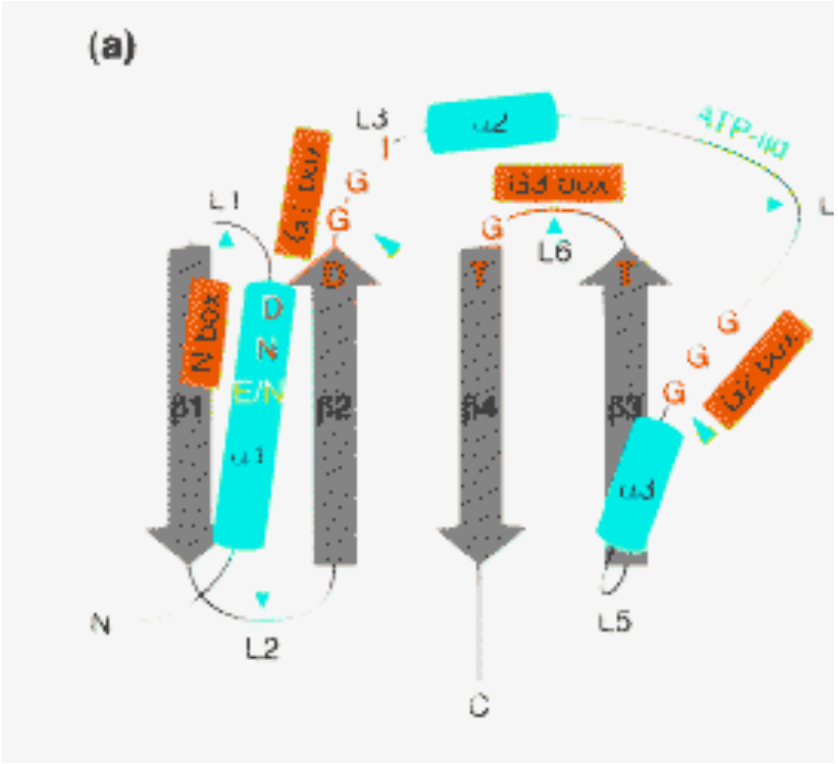
Novel ATP-binding superfamily, includes

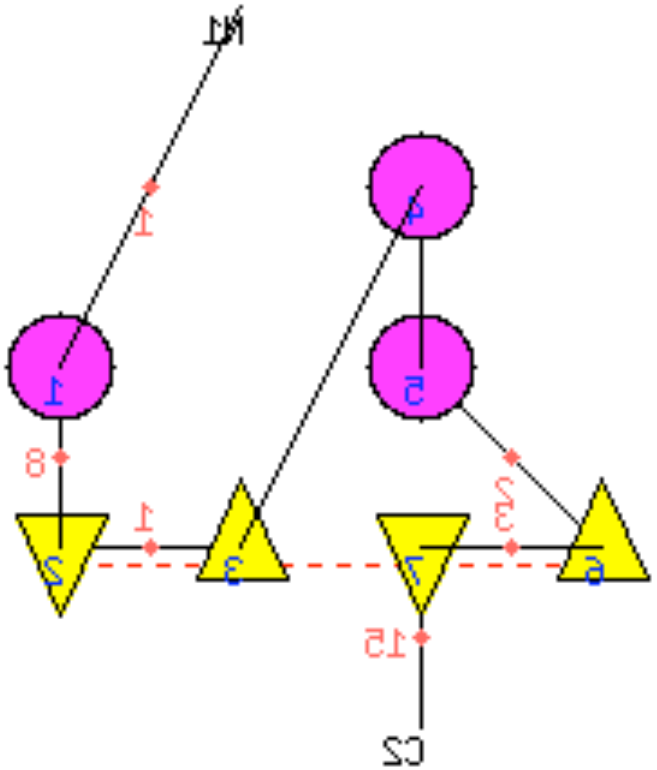
- DNA topoisomerase II
- molecular chaperones Hsp90
- DNA-mismatch-repair enzymes MutL
- histidine kinases.

The most singular unifying feature - unconventional Bergerat ATP-binding fold.

The far-reaching significance of this commonality is still in the process of being explored.

Common GHKL motif





“Patterns, functions and structures on a protein topology database”

- TOPS project, joint with Leeds (Biochem) (UK Gov’t funded)
- Development of relational database
- Addition of
  - SSE spatial neighbour info, helix packing classes & chirality relationships
  - Amino-acid sequence
  - ligand binding information
  - EC number
- TOPS visualization methods
- Improvement of search, machine learning and structure comparison algorithms & adaptation to new structural data
- Learning algorithms to relate topology to sequence and to function

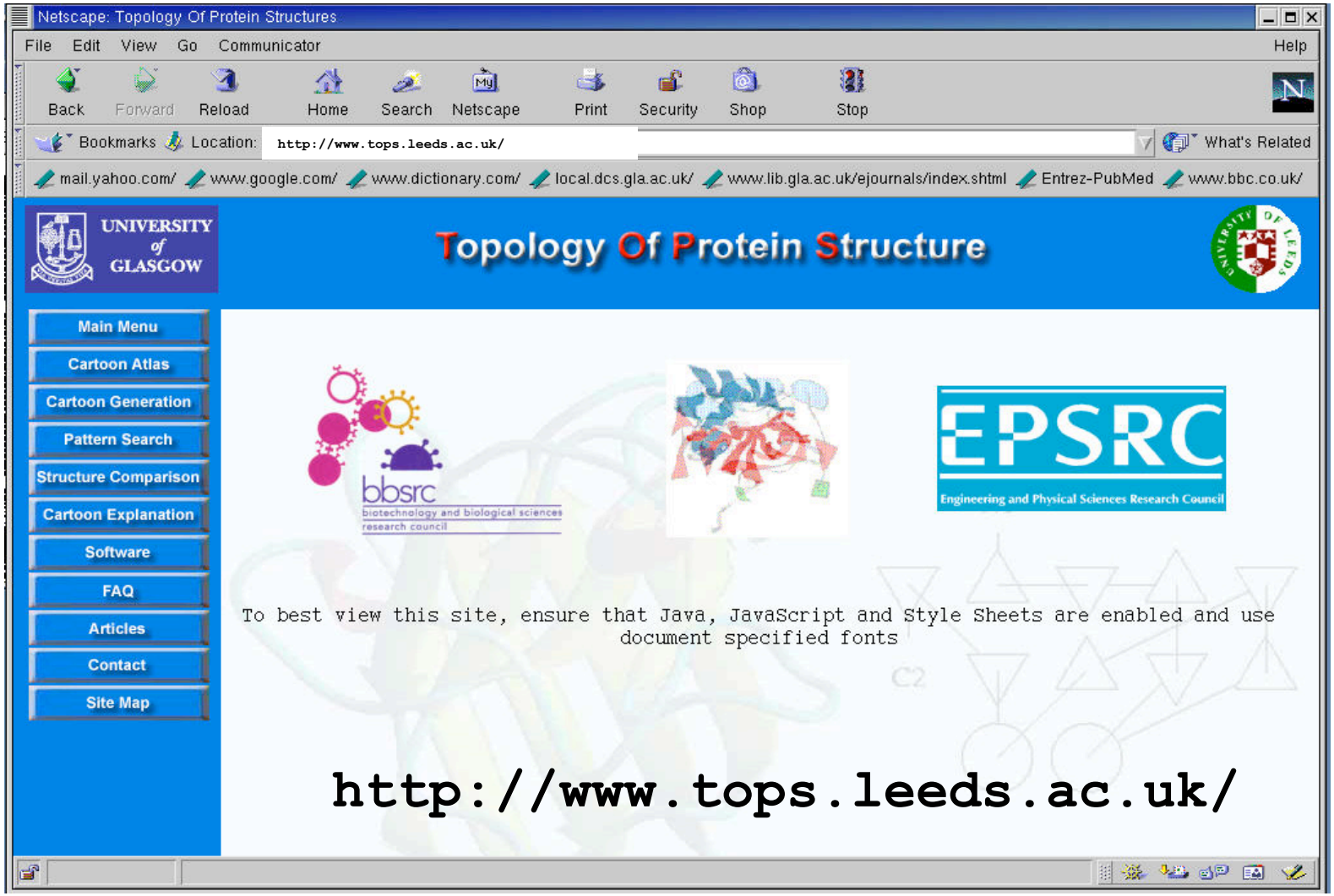
Aims

- Enhance the TOPS system, search and learning algorithms:
  - build a very fast and accurate topologically based structure search and comparison method
  - rival those based on atomic co-ordinates in terms of accuracy, excelling in terms of computational speed and memory requirements.

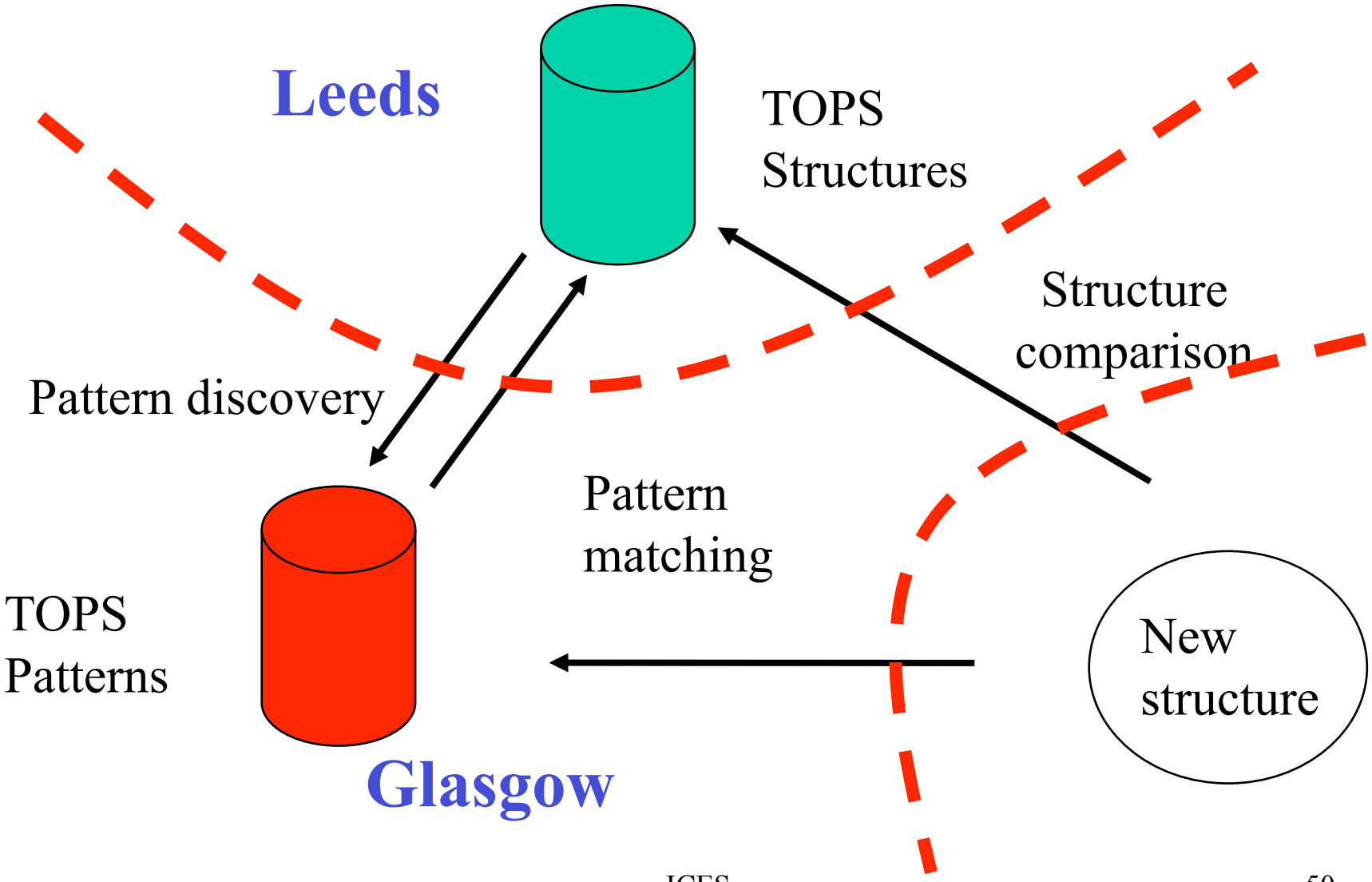
- Discover patterns relating topological descriptions to
  - sequence
  - function (e.g. ligand binding sites)

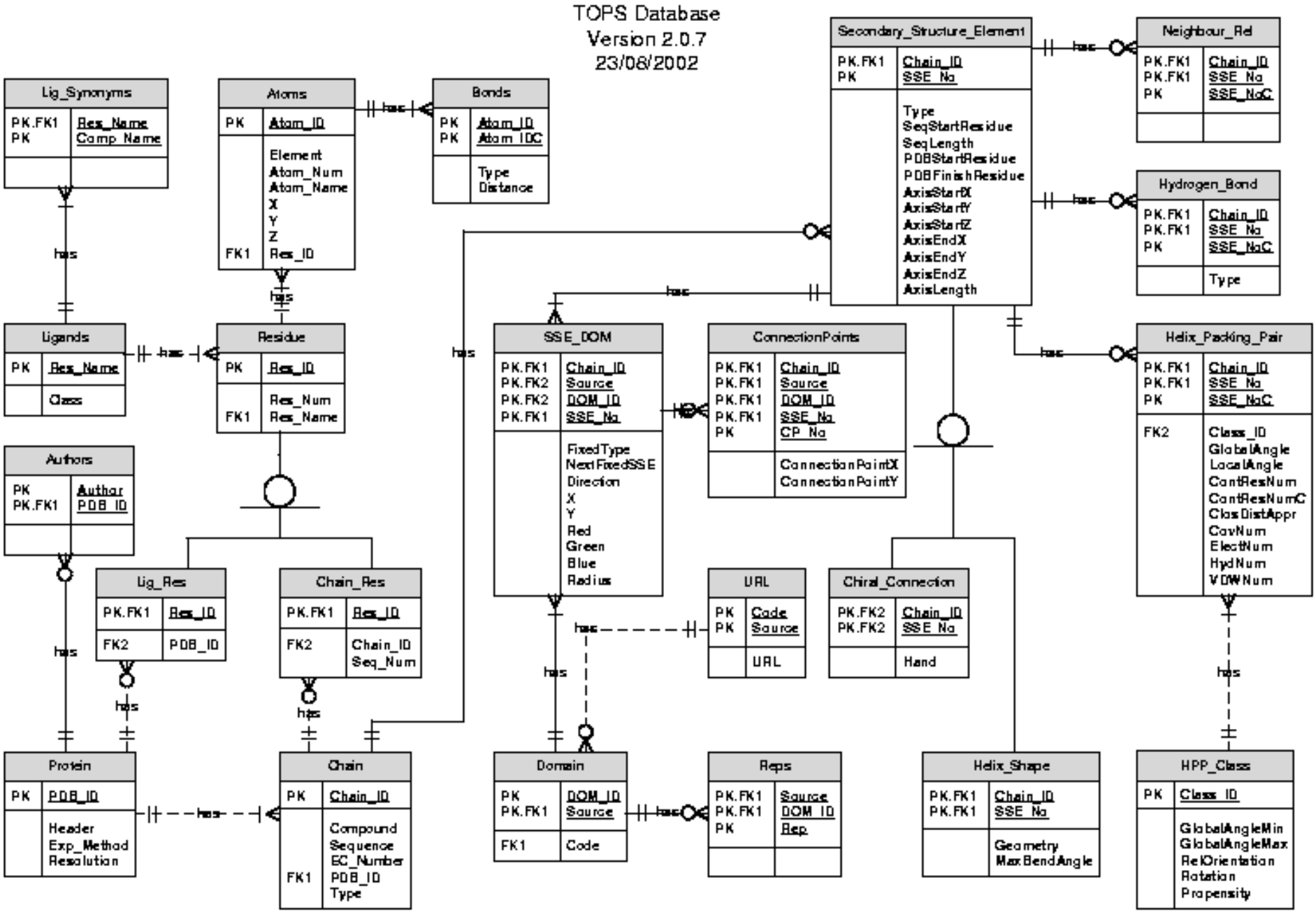
→ Use in prediction of protein structure and function.

TOPS website



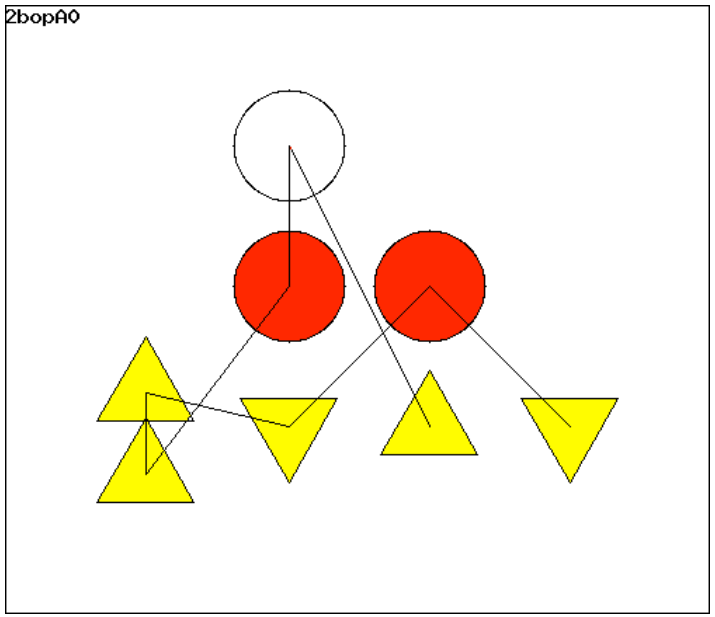
Glasgow-Leeds Project

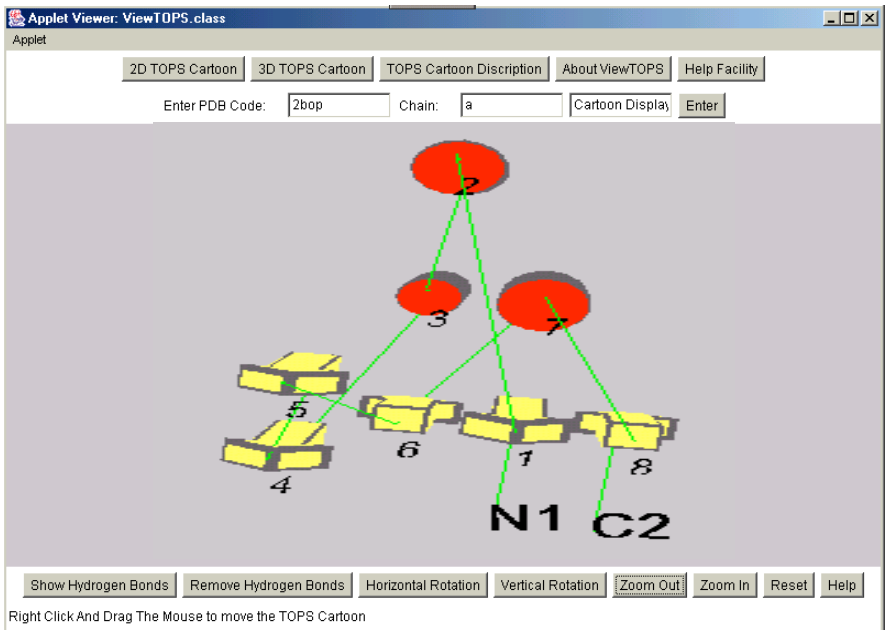


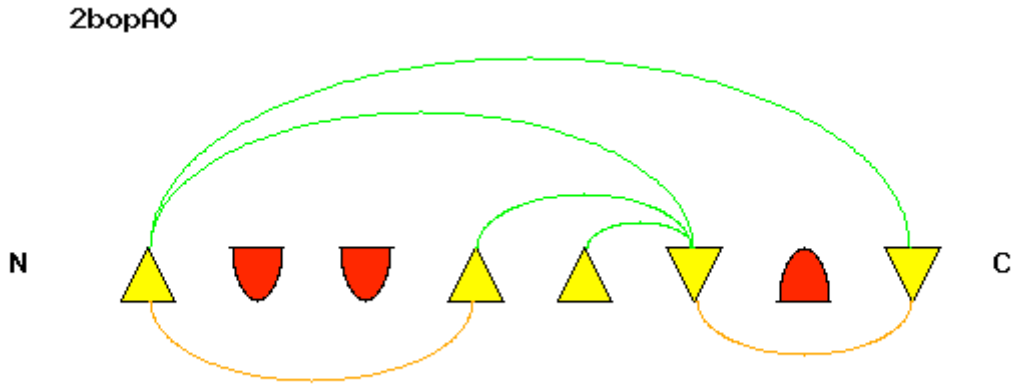

TOPS Database
Version 2.0.7
23/08/2002
Lig_Synonyms
PK.FK1 Res_Name
PK Comp_Name
Atoms
PK Atom_ID
Element
Atom_Num
Atom_Name
X
Y
Z
FK1 Res_ID
Bonds
PK Atom_ID
PK Atom_IDC
Type
Distance
Secondary_Structure_Element
PK.FK1 Chain_ID
PK SSE_No
Type
SeqStartResidue
SeqLength
PDBStartResidue
PDBFinishResidue
AxisStartX
AxisStartY
AxisStartZ
AxisEndX
AxisEndY
AxisEndZ
AxisLength
Neighbour_Rel
PK.FK1 Chain_ID
PK.FK1 SSE_No
PK SSE_NoC
Hydrogen_Bond
PK.FK1 Chain_ID
PK.FK1 SSE_No
PK SSE_NoC
Type
Ligands
PK Res_Name
Class
Residue
PK Res_ID
Res_Num
FK1 Res_Name
SSE_DOM
PK.FK1 Chain_ID
PK.FK2 Source
PK.FK2 DOM_ID
PK.FK1 SSE_No
FixedType
NextFixedSSE
Direction
X
Y
Red
Green
Blue
Radius
ConnectionPoints
PK.FK1 Chain_ID
PK.FK1 Source
PK.FK1 DOM_ID
PK.FK1 SSE_No
PK CP_No
ConnectionPointX
ConnectionPointY
Helix_Packing_Pair
PK.FK1 Chain_ID
PK.FK1 SSE_No
PK SSE_NoC
FK2 Class_ID
GlobalAngle
LocalAngle
CentResNum
CentResNumC
ClosDistAppr
CovNum
ElectNum
HydNum
VDWNum
Authors
PK Author
PK.FK1 PDB_ID
Lig_Res
PK.FK1 Res_ID
FK2 PDB_ID
Chain_Res
PK.FK1 Res_ID
FK2 Chain_ID
Seq_Num
URL
PK Code
PK Source
URL
Chiral_Connection
PK.FK2 Chain_ID
PK.FK2 SSE_No
Hand
Protein
PK PDB_ID
Header
Exp_Method
Resolution
Chain
PK Chain_ID
Compound
Sequence
EC_Number
FK1 PDB_ID
Type
Domain
PK DOM_ID
PK.FK1 Source
FK1 Code
Reps
PK.FK1 Source
PK.FK1 DOM_ID
PK Rep
Helix_Shape
PK.FK1 Chain_ID
PK.FK1 SSE_No
Geometry
MaxBendAngle
HPP_Class
PK Class_ID
GlobalAngleMin
GlobalAngleMax
RelOrientation
Rotation
Propensity
has

TOPS visualisation

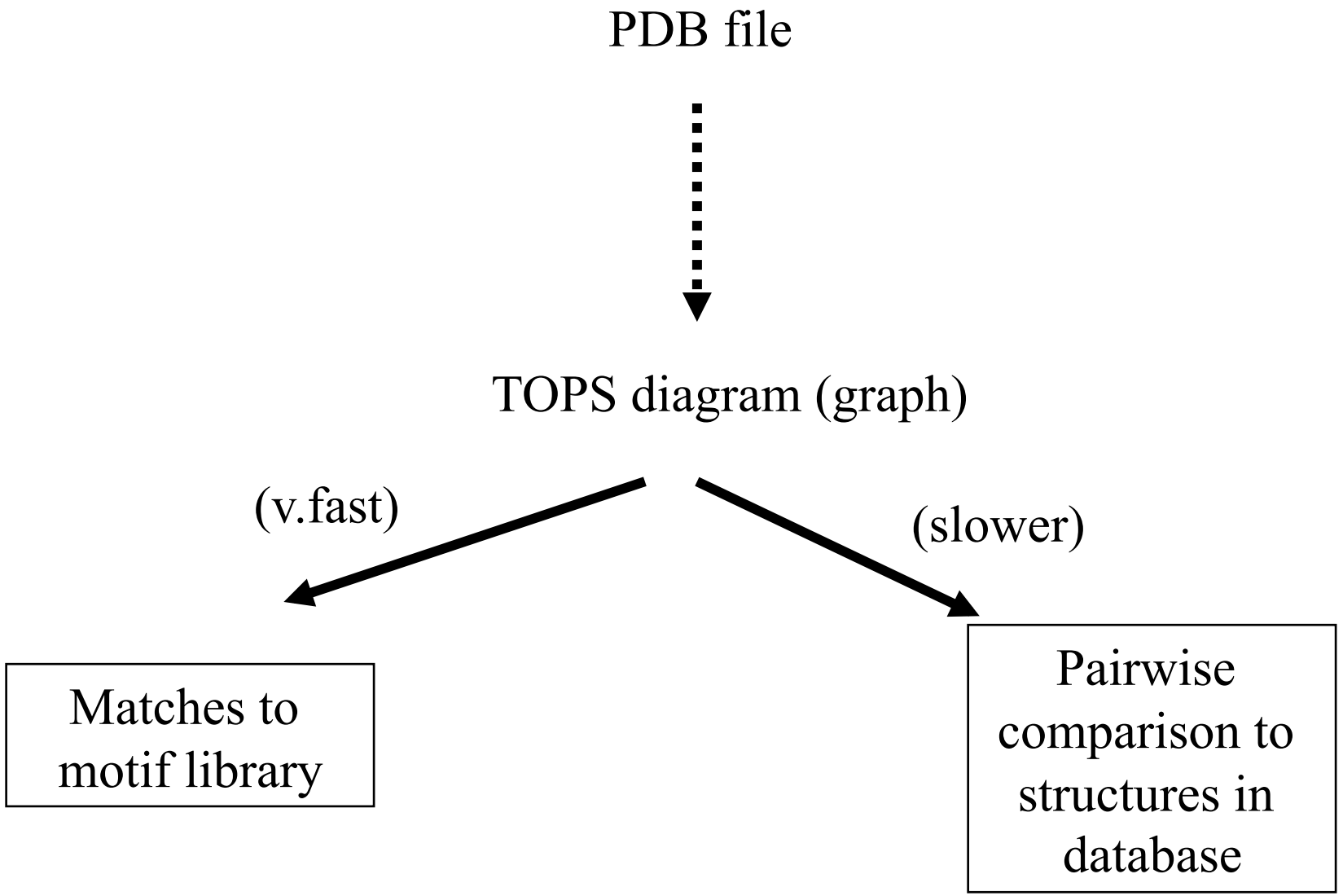
Gilleain Torrence & Neha Dhulia



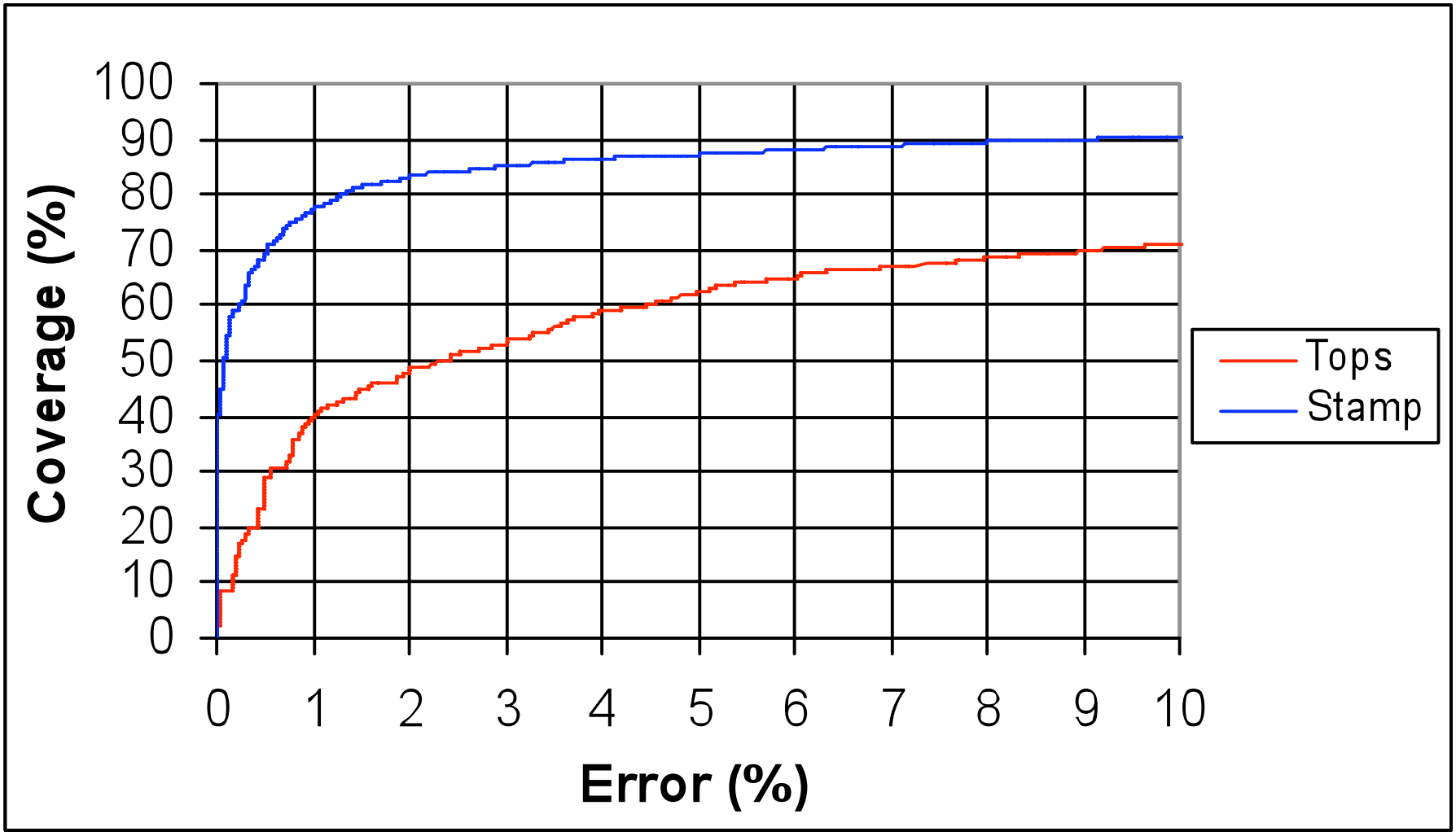




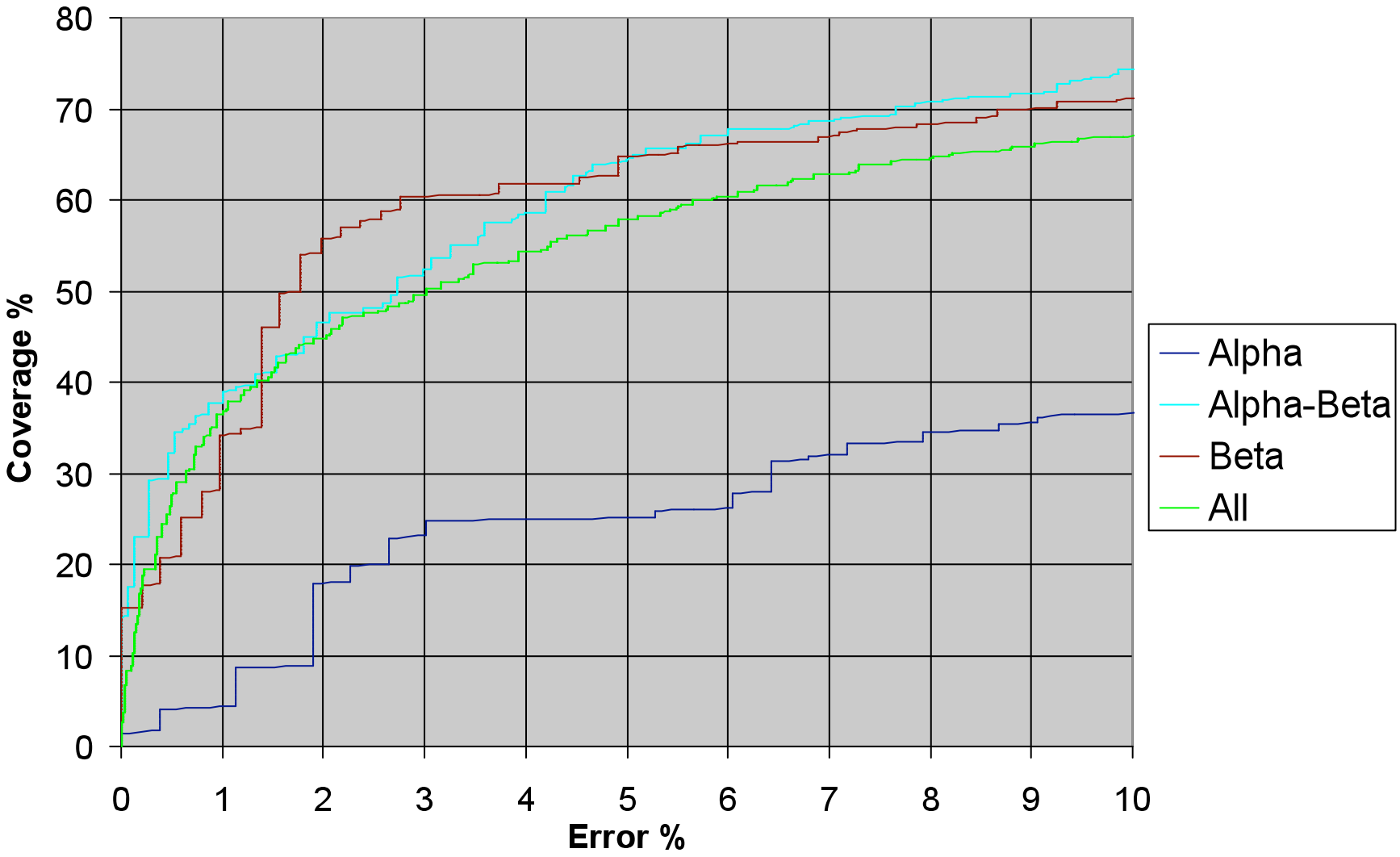
TOPS comparison server: www.tops.leeds.ac.uk

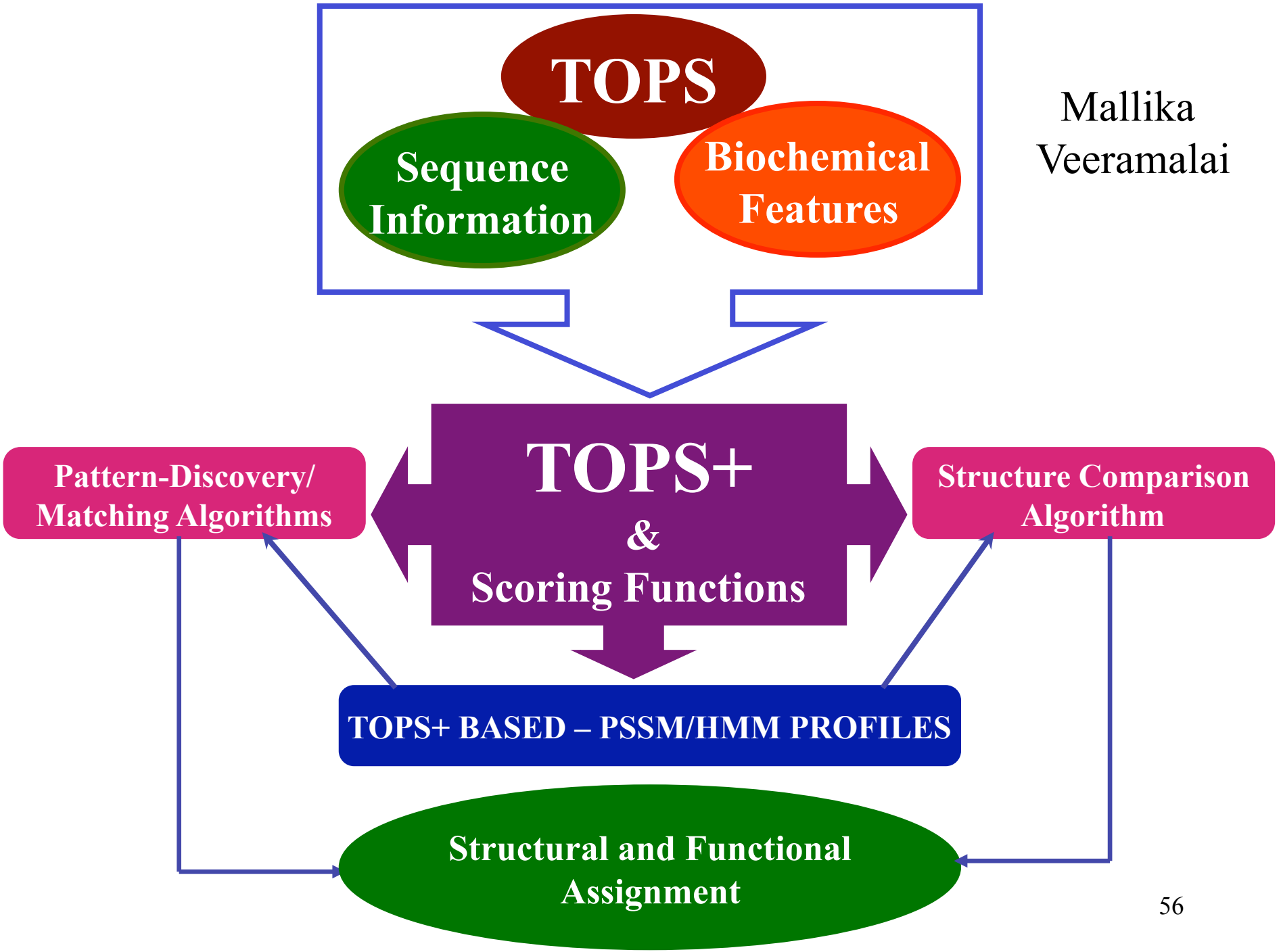


Coverage vs Error



Coverage versus error: PDB40





Mallika Veeramalai

Tops + Sequence with Biochemical Features

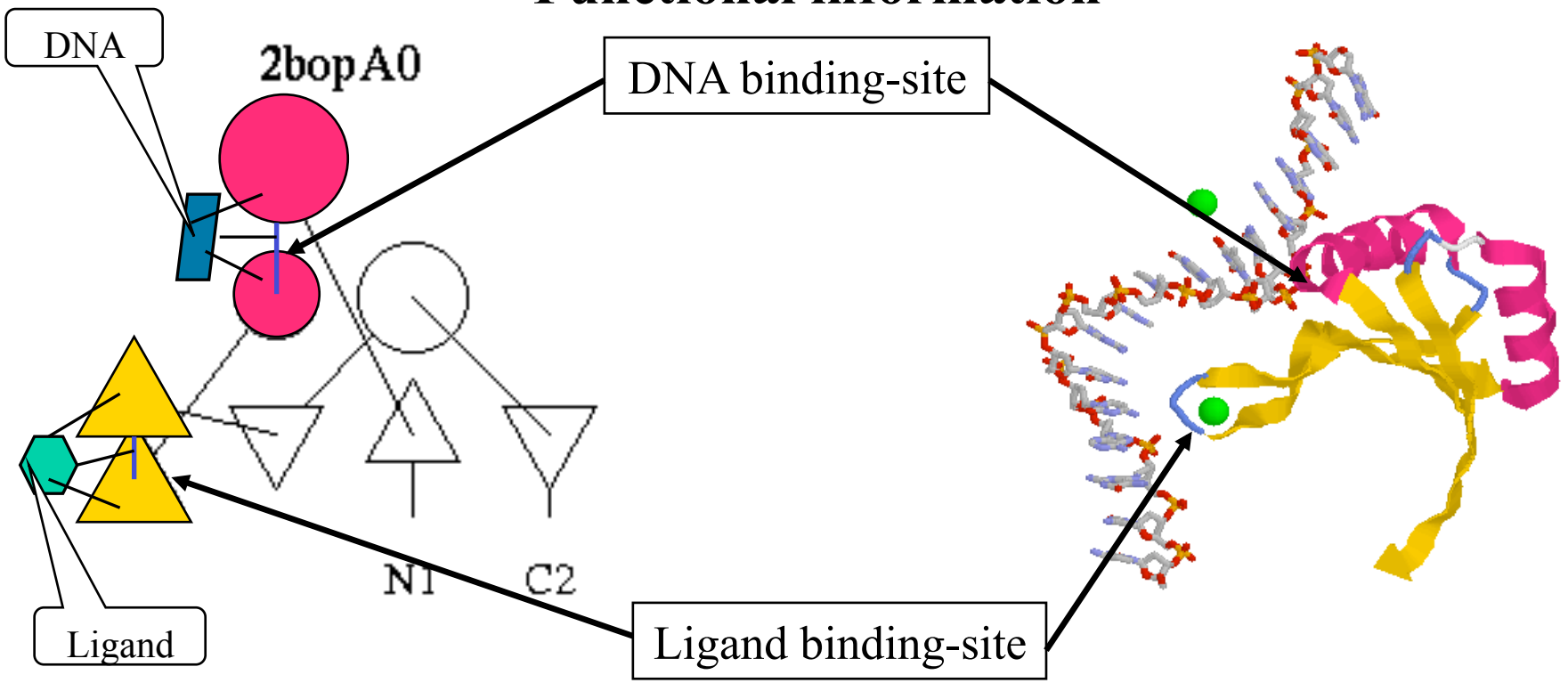
```
                          10        20        30        40        50
2bopA  ( 326 )    sCfAlIsGtangVkcYrfrVkknhrhrYencttwftvadngaerqgqAq
                   bbbbbbb aaaaaaaaaaaaaa 333  bb   bbbb      bbbbbb

                          60        70        80
2bopA  ( 376 )    IlItFgspsqrqdFlkhVplppgMnisgftasldf
                  bbbbb  aaaaaaaaaa       bbbbbb
```

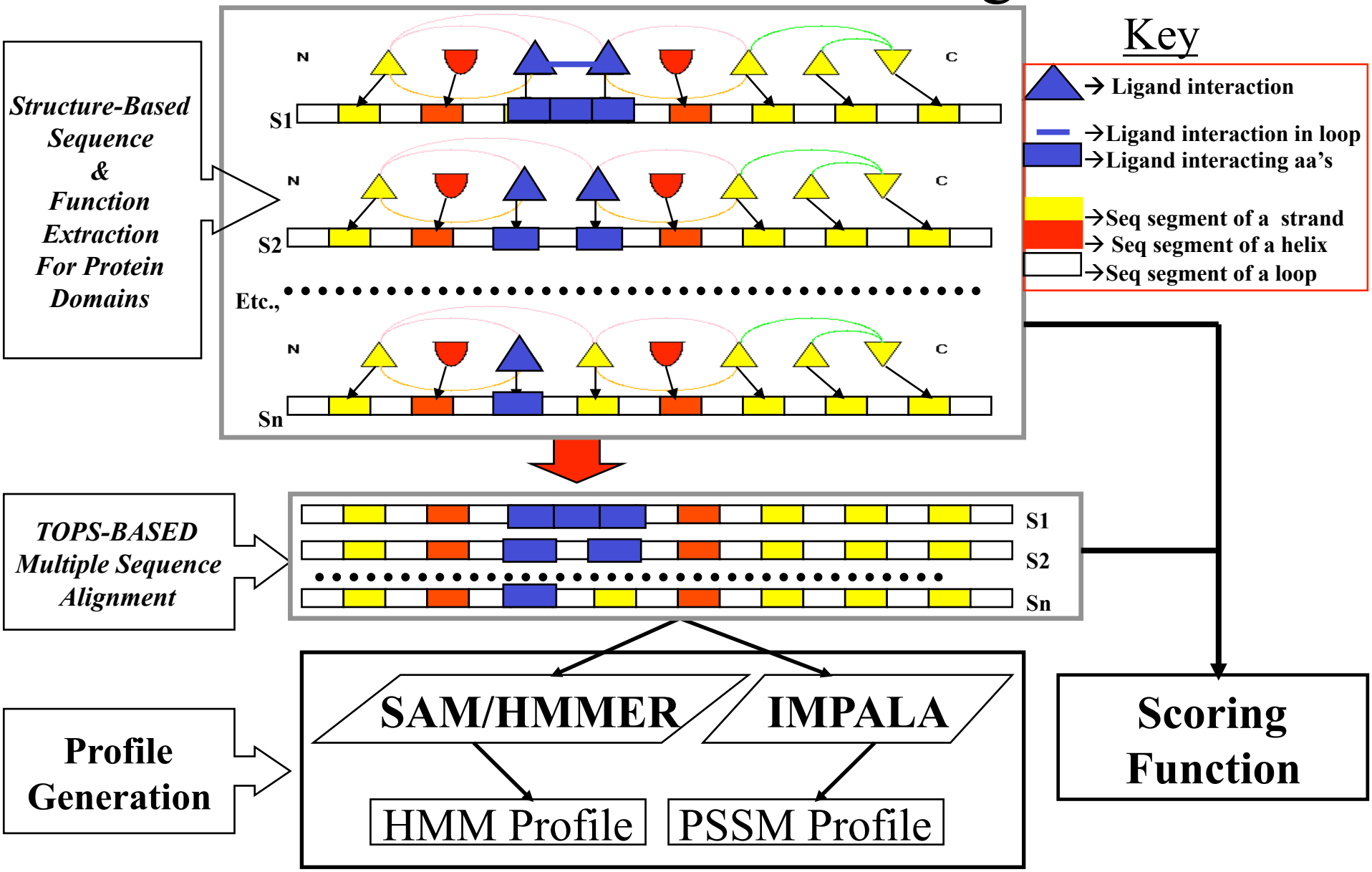
Key

solvent inaccessible	UPPER CASE	X
solvent accesible	lower case	x
alpha helix	red	x
beta strand	blue	x
3 - 10 helix	maroon	x
hydrogen bond to main chain amide	**bold**	**x**
hydrogen bond to mainchain carbonyl	underline	x
disulphide bond	cedilla	ç
positive phi	*italic*	*x*

Functional information

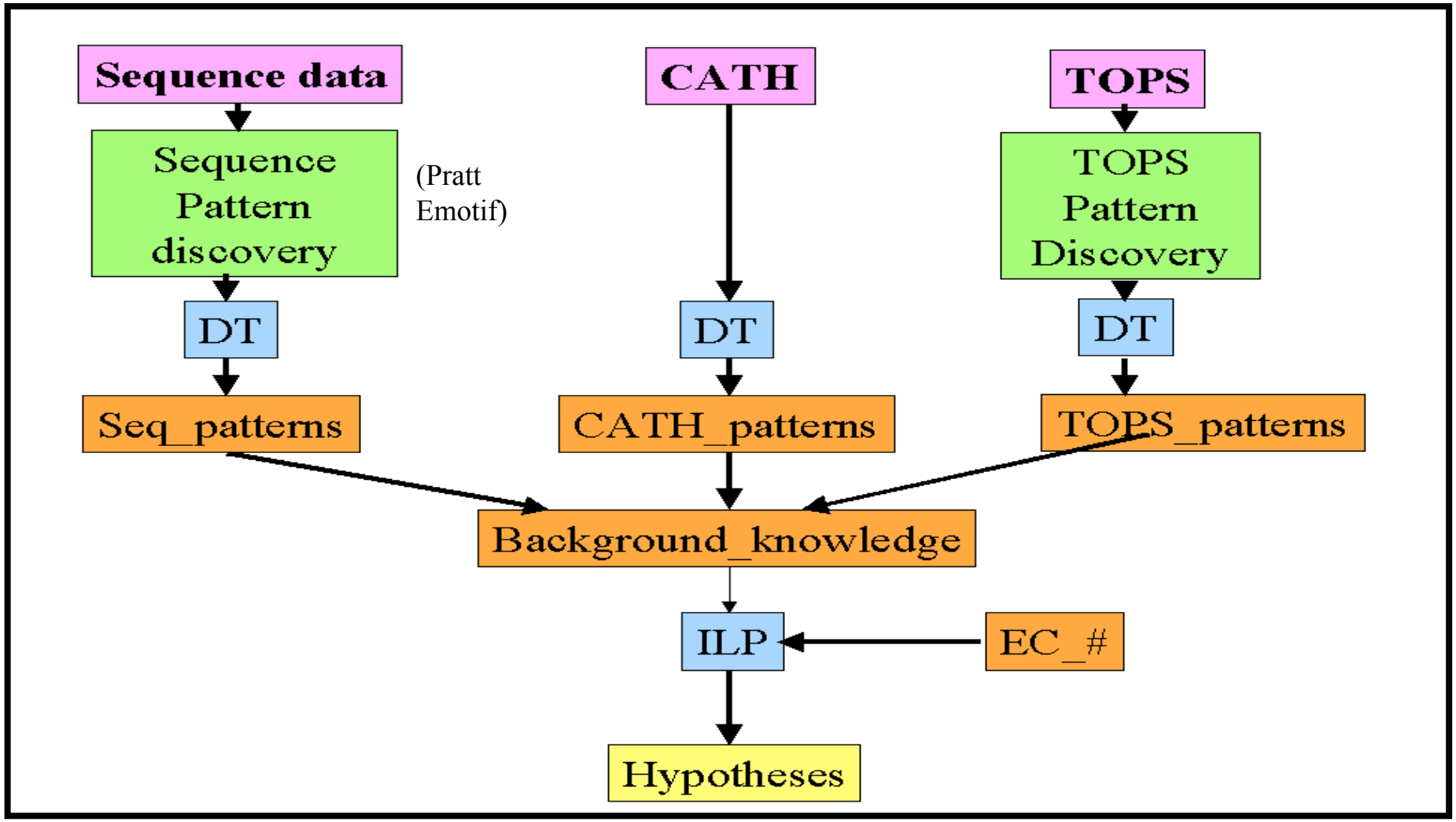


PSSM/HMM Profiles & Scoring Function



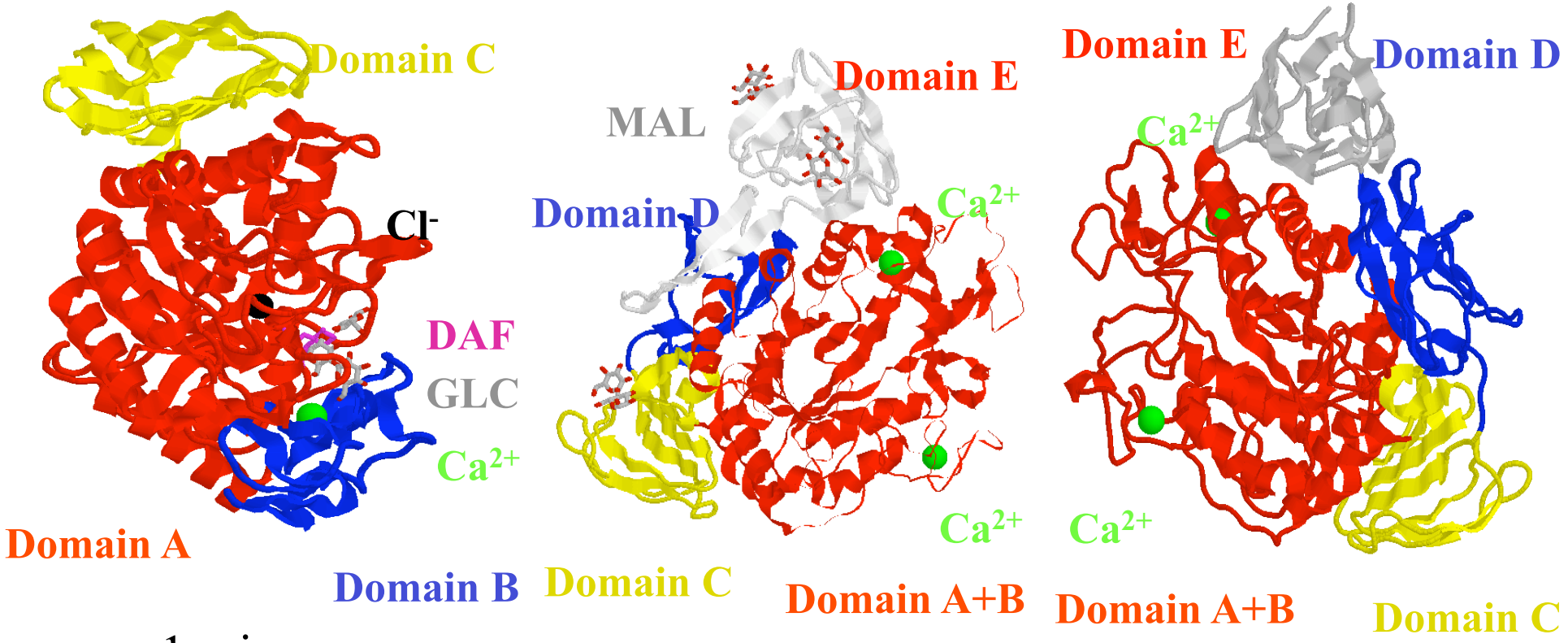
Integrative Machine Learning

Aik Choon Tan



Case Study: α-amylases Superfamily

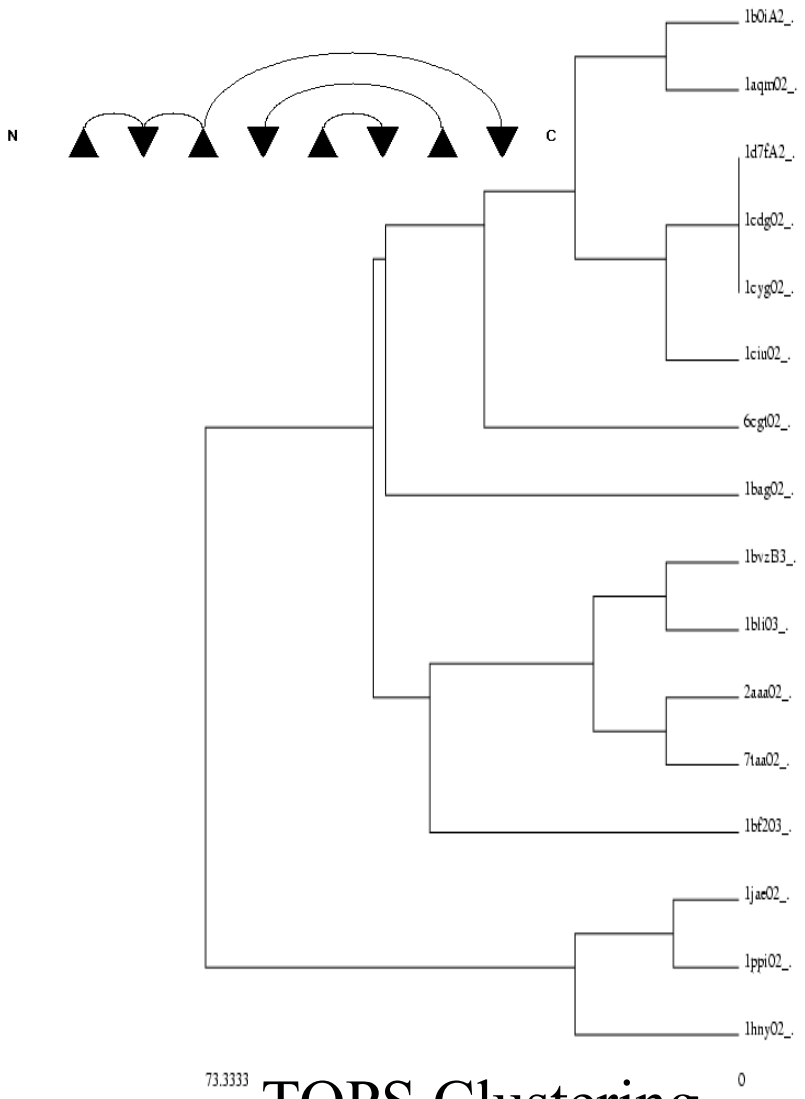
Aik Choon Tan



Domain C = Jellyrolls

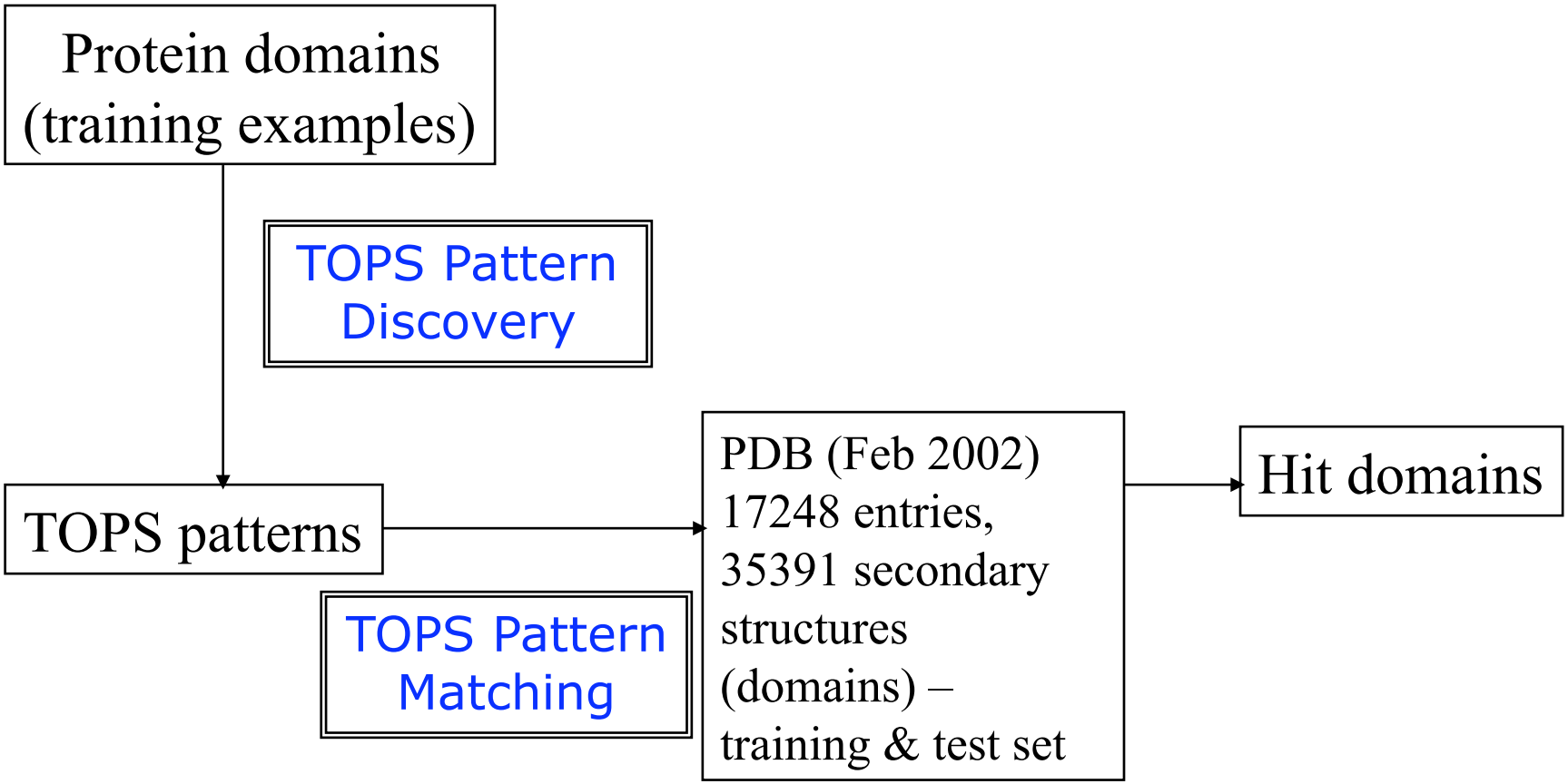
Pattern Discovery (Unsupervised)

STAMP Clustering
(3D Structure)



TOPS Clustering
(Topological)

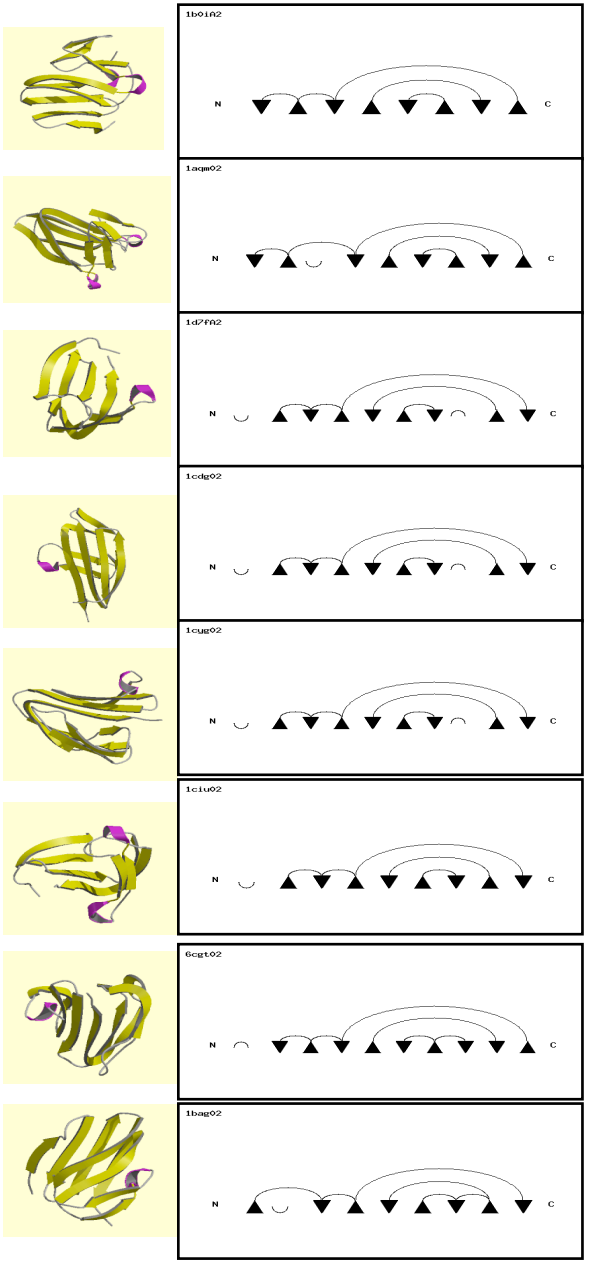
Test Set (domains)



Preliminary Results

	Pratt(TOP 5)	TOPS	IML (structures)	IML (domains)	missed (str)	discovered (str)
C 1	51	258	38	43		
C 2	608	26	9	9		
C 3	28	40	11	11		
total hits	687	324	**58**	**63**	**9**	**1**

Predictive accuracy(domains) = 63/73 = 0.86

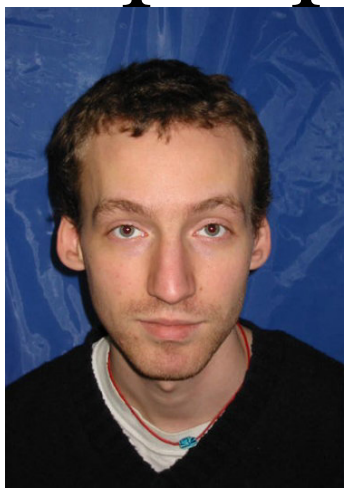


“Knowledge pattern”


```
class(A,cluster1) :-
not(has_seq_NxxNQxAFxRGxxGFxxF(A)),
not(has_seq_NxxNAxxF(A)),
ec_no(A,ec2_4_1_19), structure(A,jelly_rolls),
cath_no(A,c2_60_120_210_3), tops_pattern(A,B),
has_seq_TxLPxGxY(A),pattern(B,pattern1),
has_strands(B,8), has_hbond(B,5),
hbond_relationships(B,C,s1,D,s2,anti_parallel),
hbond_relationships(B,C,s2,D,s3,anti_parallel),
hbond_relationships(B,C,s3,D,s8,anti_parallel),
hbond_relationships(B,C,s4,D,s7,anti_parallel),
hbond_relationships(B,C,s5,D,s6,anti_parallel).
```

TOPS people

David Westhead



Gilleain Torrance

Ioannis Michalopoulos

Mallika Veeramali

Aik Choon Tan

Juris Viksna

Summary

- Formalisation of TOPS descriptions - “diagrams”
- Database of TOPS diagrams
- Motif based search facilities
- Algorithm for learning common structural patterns in a set of instances
- Application to structure comparison
- Method to cluster domains via pattern discovery

Resources / contacts

- Web sites:
  - http://tops.ebi.ac.uk/tops
  - http://www.tops.leeds.ac.uk
- Papers
  - Westhead *et al.* 1999 *Prot. Sci.* **8**:897-904
  - Gilbert *et al.* 1999 *Bioinformatics* **15**:317-326
  - Viksna & Gilbert, 2001 LNCS **2149:** 98-111

- {drg,juris,maclean}@brc.dcs.gla.ac.uk
- www.brc.dcs.gla.ac.uk

Biochemical Pathway Simulator

A Software Tool for Simulation & Analysis of Biochemical Networks

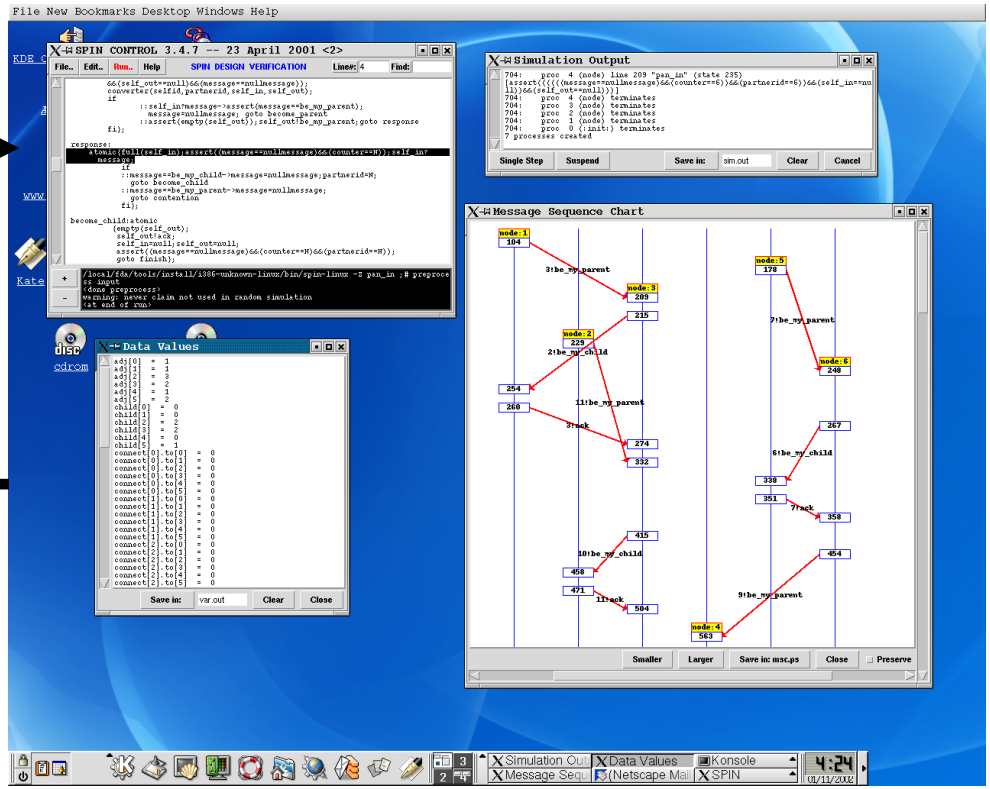
DTI 'Beacon' project, £0.9M, 4 years

Muffy Calder **David Gilbert**

Walter Kolch **Keith van Rijsbergen**

Brian Ross **Oliver Sturm**

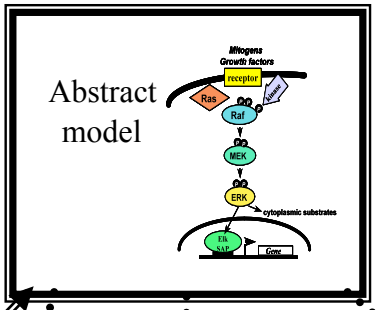
Not a toy problem!



Experimental Data

Analysis

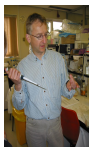
Complexity: real bioinformatics Closing the loop from wet lab to *in-silico*



Human feedback (in-the-loop)

Simulator

Lab MAPK



Database MAPK

Pathway Editor

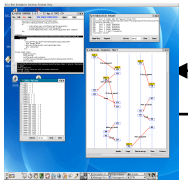
Literature Apoptosis

Text miner

Database Apoptosis

DATA

Analysis



Rules

User Interface

Web portal

Bio
Lab/Literature

Bioinformatics
Tools, database, interface

Simulator
Concurrency theory

Bioinformatics Research Centre

- Environment for collaborative interdisciplinary research in Bioinformatics.
- Hosts researchers from
  - Department of Computing Science
  - Institute of Biomedical and Life Sciences.
- Physically located in the Institute of Biomedical and Life Sciences (Davidson Building - Biochemistry & Molecular Biology)
- Strong links with
  - Sir Henry Welcome Functional Genomics Facility.
  - Statistical Bioinformatics
  - Mathematical Biology
  - NeSC Hub Glasgow
  - Protein Crystallography
- Outreach programme (visitors etc - you!)

BRC Members

- Investigators:
  - *Yves Deville* (Biochemical Networks) ***(dcs)***
  - *David Gilbert* (ML, Biochemical networks, protein structure) ***dcs***
  - Pawel Herzyk (Protein structure) ***ibls***
  - Ela Hunt (Database indexing, Data integration,...) ***dcs***
  - David Leader (Visualisation tools) ***ibls***
  - Gerhard May (Signalling pathways) ***ibls***
  - Rod Page (Phylogenetic trees) ***ibls***
  - *Olivier Sand* (Transcriptional regulatory regions) ***dcs***
  - Richard Sinnott (Grid computing / eScience) ***dcs***
  - *Juris Viksna* (Graph algorithms) ***(dcs)***
- Research Assistants: *Rainer Breitling, Neil Hanlon, Nigel Harding, Brian Ross, Oliver Sturm, Gilleain Torrance*
- Research students: *Ali Al-Shahib*, Iain Darroch, Susan Fairley, Eilidh Grant, Andrew Jones, *(Sebastian Oehm), Aik Choon Tan*, Tim Troup, *Mallika Veeramalai*
- Executive Assistant: *Fiona McBeth*
- Associated: Malcolm Atkinson, Ernst Wit, John McClure

The Scottish Bioinformatics Forum (SBF)

- Network of Bioinformatics researchers and industries in Scotland
- A vehicle for developing Scotland as a Centre of Bioinformatics Excellence
- Nodes in Glasgow, Edinburgh, Dundee, Aberdeen, ...
- Promoting collaborative research
- Development of a Bioinformatics educational programme
- www.sbforum.org, sbforum-general@sbforum.org


ISMB/ECCB 2004, A joint meeting of the...
12th International Conference on Intelligent System for Molecular Biology (ISMB 2004)
-and the-
3rd European Conference on Computational Biology (ECCB 2004)
-in conjunction with- Genes, Proteins and Computers VIII (GPCVIII)
ISMB/ECCB 2004 will be held July 31 - August 5, 2004
Glasgow, Scotland, UK
at the Scottish Exhibition and Conference Centre
Organized in association with: Collaborative Computational Project 11 (CCP11), European Bioinformatics, Institute (EMBL - EBI), Scottish Bioinformatics Forum, University of Glasgow
ISMB is sponsored by the International Society for Computational Biology (ISCB).
ECCB is the annual European Conference on Computational Biology.
ISCB
To watch for updates as they become available, visit www.iscb.org/ismbeccb2004.
www.iscb.org/ismbeccb2004

The Future

Closing the loop from wet lab to *in silico !*

Collaboration!

http://www.brc.dcs.gla.ac.uk